J. Intel Connect Veh

Research Article

AdaFusionNet: Adaptive Trajectory Prediction of Group Two-wheelers in Interactive Conflict Scenarios for Autonomous Vehicles

Lan Yang[1], Yuanhao Liu[1] ,Guangyue Qu[1,✉], Yueyang Wang[2], Shan Fang[1], Shoucai Jing[1], Xiangmo Zhao[1]

Cite this article: https://doi.org/10.26599/JICV.2026.9210102

ABSTRACT: The accurate prediction of two-wheeler motion trajectories is essential for ensuring the safety of autonomous vehicles (AVs), particularly at crowded intersections in China. Existing trajectory prediction approaches face limitations in representing the diverse motion patterns, dynamic constraints, and group-level behavioral characteristics of two-wheeler riders. This paper proposes an adaptive trajectory prediction framework for two-wheelers by integrating the revised social force model (RSFM) and a spatiotemporal Transformer (STTransformer) module. The RSFM models group-level interactions among two-wheelers, capturing self-organization, leader-follower dynamics, and herding behaviors. The STTransformer extracts temporal dependencies and spatial interaction patterns using multi-head spatiotemporal attention mechanisms to capture group motion patterns. A gated recurrent unit (GRU) serves as the fusion module, recurrently integrating the trajectory predictions generated by the RSFM and STTransformer together with a local traffic-scene complexity representation constructed from the numbers and mean speeds of nearby pedestrians and vehicles. Experimental results on the Signalized INtersection Dataset (SIND) show that AdaFusionNet achieves an accuracy improvement of over 11% at the 3 s prediction horizon compared with the selected baselines. The error distributions further indicate fewer large displacement errors, while the leave-one-intersection-out evaluation provides additional evidence of cross-intersection transferability, with AdaFusionNet maintaining competitive prediction performance when applied to intersections excluded from training.

KEYWORDS: Autonomous vehicles, group two-wheelers, trajectory prediction, spatiotemporal fusion, social force model.

Just Accepted

1 Introduction

Trajectory prediction plays a crucial role in ensuring the safety and reliability of autonomous vehicles (AVs), as it directly supports the decision-making subsystem. Handling infrequent safety-critical events remains a challenge for end-to-end autonomous driving systems (Li et al., 2026). The increasing number of two-wheelers in recent years has significantly increased traffic complexity. Due to the absence of explicit regulations, two-wheelers exhibit high randomness and uncertainty. At intersections, the operation space for different road users is often ambiguous and shared by both vehicles and two-wheelers. Behaviors such as running red lights, moving against traffic, or encroaching into motor vehicle lanes frequently occur, making their trajectories more difficult to predict and posing greater safety risks to AVs (Huang et al., 2021).

The trajectories of two-wheelers are affected by various factors, including rider intentions, road geometry, surrounding riders, vehicles, and pedestrians. Their riding behaviors often exhibit strong group characteristics; for example, a leading two-wheeler tends to assert its right of way, while others in the group closely follow. Consequently, the primary challenge in trajectory prediction for two-wheelers lies in capturing their high degree of uncertainty and complex interactive relationships to enhance prediction accuracy.

In recent years, extensive research has been conducted in the field of trajectory prediction, with most studies focusing on vehicle trajectories, while limited attention has been paid to two-wheelers. Existing trajectory prediction approaches are generally classified into three categories: physics-based methods, deep learning-based methods, and fusion methods. Physics-based approaches, such as Monte Carlo (Wang et al., 2019), Kalman filtering (Li and Jiao, 2024), and social force models (SFMs) (Liu et al., 2024, 2022), employ physical models for short-term trajectory prediction. These methods provide good interpretability but rely on simplified assumptions and exhibit limitations in capturing long-term temporal dependencies. Deep learning-based models, including generative adversarial networks (GANs) (Wei et al., 2024), long short-term memory (LSTM) (Feng et al., 2025), Transformers (Chen et al., 2025; Xu et al., 2025), and graph neural networks (Zhou et al., 2021), demonstrate strong capability in learning complex patterns and long-term dependencies from data. Currently, an increasing number of AV trials adopt deep learning-based methods to predict the future trajectories of surrounding traffic participants. However, these models often fail to adequately integrate physical constraints, resulting in predictions that violate basic kinematic principles and exhibit limited interpretability. To overcome the limitations of the aforementioned models, fusion approaches have been proposed. These models are generally classified into pixel-level, feature-level, and decision-level categories, corresponding to the front-end, mid-end, and back-end stages, respectively. However, fusion

[1] School of Information Engineering, Chang'an University, Xi'an 710018, China. [2] Institute for Transport Studies, University of Leeds, Leeds LS2 9JT, UK.

✉ Corresponding author. E-mail: 2023224092@chd.edu.cn

Received: July 12, 2026; Revised: August 10, 2026; Accepted: September 19, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/).

models often encounter challenges such as limited scalability, difficulties in feature alignment, and redundant fusion dimensions when handling game-theoretic interactions (Rudenko et al., 2018).

The aforementioned methods generally predict targets as isolated entities or consider the surrounding environment as independent, neglecting the social interaction characteristics and behavioral causality within two-wheeler groups, which leads to reduced prediction accuracy. Therefore, this study proposes an adaptive trajectory prediction approach for group two-wheelers in interactive conflict scenarios. Although social force models, Transformers, and recurrent neural networks are established techniques, the methodological contribution of this study lies in their task-specific coordination for group two-wheeler prediction. The framework assigns complementary functions to an interpretable spatial-interaction branch and a data-driven spatiotemporal branch, converts their outputs into a common trajectory space, and performs scene-complexity-informed recurrent fusion over the prediction sequence. The main contributions are summarized as follows:

1) A two-wheeler group interaction model based on the revised social force model (RSFM) is proposed, which accounts for the continuous interactions between two-wheelers and surrounding objects, while also incorporating group herding behaviors and intra-group cooperative-competitive relationships.

2) A deep learning model based on the Spatiotemporal Transformer (STTransformer) is proposed, which employs a cross-attention mechanism to capture the long-term temporal dependencies of two-wheeler groups, enabling accurate modeling of their overall motion trends and internal cooperative–competitive behaviors.

3)A GRU-based multi-model trajectory fusion method is introduced to integrate the complementary trajectory predictions generated by the RSFM and STTransformer. The fusion mechanism incorporates the scene complexity representation $E(\mathbf{C})$ and performs recurrent temporal fusion over the prediction sequence, enabling more accurate trajectory prediction for two-wheeler groups.

2 Related work

Current research on trajectory prediction models can be classified into three categories: physics-based models, deep learning-based models, and fusion models.

2.1. Physics-Based Trajectory Prediction

Physics-based trajectory prediction for two-wheelers relies on predefined dynamic and kinematic models to simulate potential motion paths. These approaches employ physical laws and mechanistic assumptions to model individual movements and interactions. Among them, the social force model (SFM) is widely used because of its interpretability and its ability to represent interaction-driven motion through force-based formulations.

The effectiveness of SFM-based methods has been demonstrated in several two-wheeler and mixed-traffic trajectory prediction studies. Enhanced SFM formulations have been applied to cyclist trajectory prediction at unsignalized intersections, where model parameters were calibrated using maximum likelihood estimation to better capture cyclist crossing behavior in heterogeneous traffic (Huang et al., 2016). Subsequent studies further extended SFM to intersection conflict scenarios by considering multi-participant interactions and optimizing model parameters through maximum likelihood estimation (Zhu et al., 2019). For pedestrian—electric bicycle mixed flows at crosswalks, modified SFM variants have incorporated bicycle geometry, visual field range, and avoidance behavior to improve interaction modeling (Fu et al., 2023). In bidirectional pedestrian--two-wheeler mixed flows, additional boundary forces, overtaking forces, and following behavior forces have been introduced to better describe heterogeneous interactions, with model parameters calibrated using genetic algorithms (Wang et al., 2024; Zhang et al., 2024). Although these physics-based models exhibit strong extensibility and can flexibly incorporate environmental factors, they are mainly effective for short-term prediction and tend to suffer from rapidly increasing errors over longer horizons due to their dependence on instantaneous information and the lack of temporal dependency modeling (Li et al., 2023a).

2.2. Deep Learning-based Trajectory Prediction

To address the challenge of interaction modeling for two-wheelers, deep learning-based trajectory prediction models have been proposed. These models utilize large-scale datasets to learn the complex motion patterns of two-wheelers. Image-based learning frameworks have been used to extract features from image and optical-flow sequences and combine them with historical trajectory features for cyclist position prediction, demonstrating strong applicability in real traffic environments (Zernetsch et al., 2021). Spatial-temporal multi-graph structures have also been introduced to model social interactions, while multi-kernel functions and temporal convolution have been used to represent social attributes and enhance prediction performance in dense traffic scenarios (Li et al., 2022). Hierarchical convolutional networks combined with multi-task learning have further been adopted to capture heterogeneous behavior patterns and support unified multi-modal trajectory prediction for cyclists around AVs (Li et al., 2023b). Multi-view adaptive hierarchical spatial graph convolutional networks have been developed for trajectory prediction in heterogeneous traffic, achieving superior performance for cyclists(Xu et al., 2023). Transformer-based frameworks have also emphasized the importance of modeling social interactions with the environment, enabling multi-class participant handling and rule-compliant trajectory generation (Li et al., 2024b). These studies highlight the substantial influence of interactions between two-wheelers and surrounding participants on trajectory evolution. However, their performance remains heavily dependent on the quality of training data, leading to poor generalization in unseen scenarios and limited capability in handling edge cases. Moreover, the internal decision-making process of these models lacks interpretability and logical traceability, preventing the capture of underlying causal behavioral mechanisms.

2.3. Fusion Model for Trajectory Prediction

To overcome the limitations of single-class models, researchers have increasingly adopted multimodal fusion approaches for trajectory prediction. Dynamic Bayesian networks have been integrated with LSTM models, where intention inference is first performed to support faster cyclist motion prediction and LSTM is then used to improve trajectory accuracy (Gao et al., 2021). Other fusion frameworks have combined improved SFM variants with attention-based recurrent networks and ensemble strategies, including AN-LSTM, IA-GRU, stacking, and AdaBoost, to improve the trajectory prediction accuracy

Just Accepted

of vulnerable road users (Chen et al., 2022, 2023). Two-layer integrated frameworks have also been developed for bicycle trajectory prediction, in which multimodal trajectories are first generated using convolutional neural networks and then selected according to the principle of maximum efficiency, thereby achieving strong performance in complex traffic scenarios (Li et al., 2024a). In contrast to serial hybrid pipelines and ensemble methods based on static weighting, AdaFusionNet retains the RSFM and STTransformer as independent predictors and performs recurrent decision-level fusion after both branches have generated temporally aligned trajectories. The scene complexity representation $\boldsymbol{E}(\mathbf{C})$ and the previous GRU state jointly participate in the fusion process, allowing the recurrent fusion state to evolve over the prediction sequence without directly coupling heterogeneous branch-internal representations.

To summarize, the key findings from the literature review are as follows:

1) Two-wheelers exhibit pronounced group characteristics, indicating that their trajectory prediction should not be performed in isolation. However, research on simultaneous trajectory prediction for two-wheeler groups remains limited.

2) Existing studies primarily focus on typical traffic scenarios, while transfer performance under unseen or different intersection conditions remains insufficiently evaluated. This limitation motivates further investigation of trajectory-prediction methods across heterogeneous traffic scenarios.

3) Existing fusion approaches mainly adopt serial hybrid structures or ensemble strategies based on relatively static combination schemes. Although these approaches can exploit complementary information from different models, recurrent decision-level fusion of independently generated and temporally aligned trajectory predictions remains insufficiently explored, particularly when local traffic-scene information is incorporated into the fusion process.

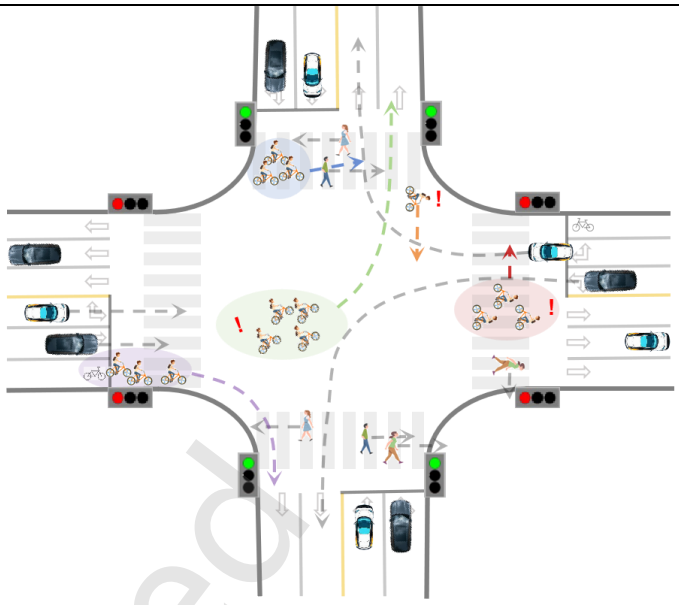


Fig. 1. Two-wheeler traffic scenarios at signalized intersections.

3 Methodology

3.1. Problem Description

Two-wheelers at intersections exhibit both group and individual behavioral characteristics (错误!未找到引用源。). From a group perspective, two-wheelers exhibit herding behavior by following preceding riders to enhance traffic efficiency or safety, occasionally even collectively violating traffic regulations, such as running red lights or moving against traffic. At the individual level, two-wheelers demonstrate high flexibility, dynamically adjusting their riding strategies and speed control based on real-time traffic conditions and available spatial resources.

In this study, a two-wheeler group is defined from the perspective of the autonomous vehicle rather than as a fixed-size platoon or a manually labeled social group. The subject vehicle refers to the reference motor vehicle associated with the current interaction scene, while an interaction sample consists of a fixed historical observation window and its subsequent prediction window. All two-wheelers retained as prediction targets within the same interaction

Just Accepted

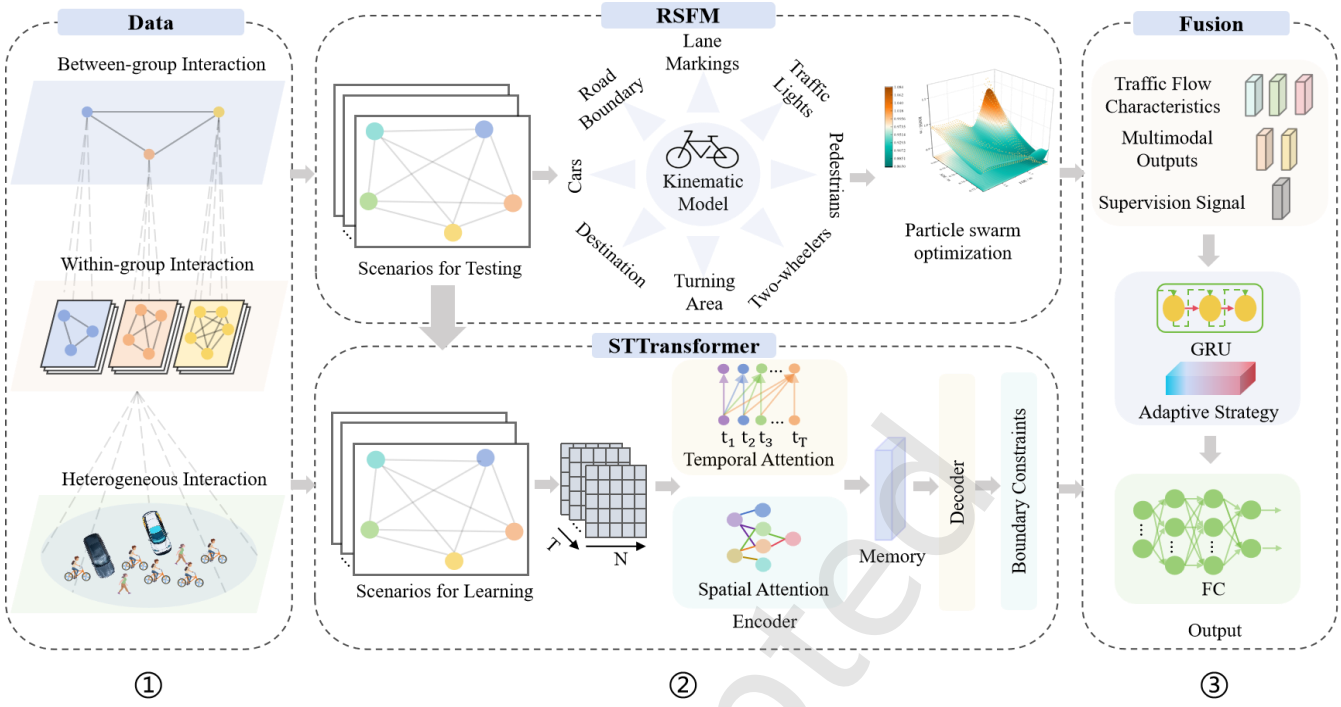


Fig. 2 Architecture of proposed trajectory prediction model for group two-wheelers.

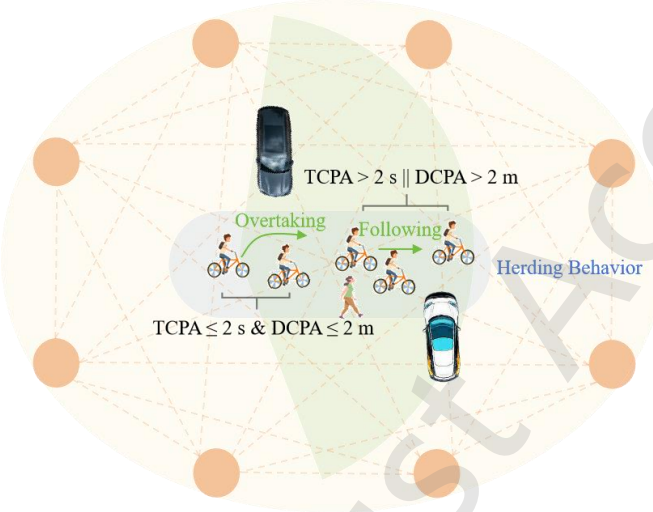


Fig. 3 Herding behaviors and intra-group cooperative-competitive relationships in two-wheeler groups.

sample constitute the scene-level two-wheeler group. The group size is determined by the actual traffic condition and may vary across samples. Cars and pedestrians in the same scene are treated as heterogeneous context agents rather than members of the two-wheeler group. The 10 m neighborhood threshold introduced later is used only for constructing the target-specific local interaction graph of each predicted group member and does not define the membership or size of the scene-level two-wheeler group.

The trajectory prediction task for two-wheeler groups is formulated as a deterministic scene-level multi-two-wheeler trajectory prediction problem. Given the observed historical trajectory sequences of the target two-wheelers $\mathbf{X}^h$ and the dynamic environmental context $\mathbf{C}$, AdaFusionNet predicts the future joint trajectory as

$$\hat{Y}^f = f_\theta(\mathbf{X}^h, \mathbf{C}), \tag{1}$$

where $\mathbf{X}^h$ denotes the observed historical states of the interacting two-wheelers, C represents the contextual information available during the observation period, including road geometry, traffic-signal states, and surrounding traffic participants, and $\hat{Y}^f$ denotes the predicted future joint trajectory. For a given historical observation and traffic context, the model produces one deterministic trajectory for each target two-wheeler rather than multiple candidate trajectories or an explicit probability distribution. The proposed framework contains two complementary trajectory-prediction branches. The first branch is the rule-based RSFM, which models group-level spatial interactions through the internal interaction representation F_t, given by

$$F_k^i = \sum_{\substack{j=1 \\ j \neq i}}^{N} \phi\left(y_{k-1}^i, y_{k-1}^j, C\right) \tag{2}$$

where $\phi(\cdot)$ denotes the interaction function that computes the influence of agent j on agent i, while F_k^i represents the aggregated interactive feature for agent i at time k. The second branch is the data-driven STTransformer, which captures spatiotemporal dependencies through the internal representation H_t, given by

$$H_k = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K and V represent the query, key, and value matrices derived from the historical joint states $X\{1:T_h\}$, respectively, allowing the model to learn the spatiotemporal patterns of the entire two-wheeler group. After completing their respective internal interaction and spatiotemporal modeling processes, the RSFM and STTransformer independently generate future trajectory predictions. Let $\hat{\mathrm{y}}_k^R$ and $\hat{\mathrm{y}}_k^T$ denote the trajectory predictions generated by the RSFM and STTransformer at prediction step k, respectively. The two trajectory outputs are temporally aligned and represented in the same planar coordinate system before entering the fusion module. Together with the scene complexity representation $E(\mathbf{C})$, they are used to update the recurrent fusion state as

Just Accepted

$$z_k = \mathrm{GRU}(z_{k-1}, \left[\hat{y}_k^R, \hat{y}_k^T, E(\mathbf{C})\right]) \quad (4)$$

Here, $E(\mathbf{C})$ denotes the local traffic-scene complexity representation constructed from the observed surrounding traffic state at the beginning of prediction. It consists of the numbers and mean speeds of nearby pedestrians and vehicles. $E(\mathbf{C})$ is concatenated with the temporally aligned trajectory predictions generated by the RSFM and STTransformer as contextual information for recurrent fusion. It does not directly determine the reset or update gates; instead, the gate values are jointly learned from the complete fusion input and the previous hidden state.

The recurrent state z_k is subsequently mapped to the fused trajectory prediction at prediction step k, and the sequence of fused predictions constitutes the final joint trajectory $\hat{Y}^f$ defined in (1).

3.2. Framework

The proposed method integrates the group spatial constraints captured by the RSFM with the long-term temporal dependency modeling capability of the STTransformer (**Fig. 2**). A GRU-based network is employed to recurrently fuse the complementary trajectory predictions generated by the RSFM and STTransformer together with the scene complexity representation $E(\mathbf{C})$,producing the final trajectory prediction for each target two-wheeler.

The first phase involves data preprocessing and environmental information construction. Stationary and anomalous trajectory points are removed from the naturalistic driving dataset, and linear interpolation is applied to ensure trajectory continuity. In the second phase, two complementary prediction models are developed in parallel. A kinematic-based RSFM is established to incorporate multiple influencing factors, including destination-driven forces, interactive forces from surrounding agents (such as vehicles, pedestrians, and other two-wheelers), steering forces in turning zones, and constraints from road geometry and traffic signals. Concurrently, an STTransformer model is implemented to capture group-level historical motion patterns. This is accomplished through spatiotemporal attention mechanisms and an adjacency matrix masking strategy, enhancing the model's capability to capture complex temporal dependencies and structured spatial interactions within the two-wheeler group. The third phase focuses on implementing an adaptive network layer for multi-model fusion. A two-layer GRU is employed to integrate inputs from three dimensions: trajectory forecasts of two-wheeler groups generated by the RSFM, those produced by the STTransformer, and the state information of the surrounding traffic flow. To this end, the two-layer GRU integrates the trajectory predictions generated by the RSFM and STTransformer together with the scene complexity representation $E(\mathbf{C})$. Future ground-truth trajectories are used as supervision during training. Through recurrent hidden-state updates, the fusion module learns a nonlinear temporal mapping from the complementary trajectory predictions and scene complexity representation to the final predicted trajectories of the two-wheeler group.

3.3. RSFM Model Considering Group Spatial Constraints

The trajectories of two-wheelers at intersections are influenced by multiple factors, including destination attraction, group interactions with surrounding traffic participants, turning intentions, road geometry constraints, and traffic signal phases. These forces collectively govern the complex motion patterns exhibited by two-wheelers.

In the RSFM, five social forces are considered: the self-motivation force F_a , the dynamic interaction force F_{ij} (arising from interactions with pedestrians, two-wheelers, and vehicles), the road boundary constraint force F_b and F_l , and the decision-region steering force F_t at intersection entry zones. The resultant force F_{ic} generated by these components is given by

$$F_{ic} = F_d + F_{ij} + F_b + F_l + F_t \quad (5)$$

3.3.1. Self-Motivation Force

From the historical trajectory vectors P_s and P_e of two-wheelers, the unit vector P_o representing the desired direction is computed. The destination-directed self-motivation force F_d is then derived as follows:

$$e_o = \frac{p_e - p_s}{\|p_e - p_s\|} \quad (6)$$

$$F_d = \frac{v_d e_o - v_i}{\tau_i} \quad (7)$$

where v_d denotes the desired speed magnitude, v_i represents the actual velocity vector of two-wheeler i ,and τ_i indicates the relaxation time of two-wheeler i .

3.3.2. Dynamic Group Interaction Force

Groups of two-wheelers at intersections often exhibit distinct herding behaviors, where individuals follow collective movements to enhance safety and efficiency (**Fig. 3**). However, underlying this group cohesion are simultaneous competitive and cooperative dynamics, as two-wheelers engage in subtle interactions to negotiate spatial resources.

Two-wheelers are influenced by the collective motion of preceding vehicles, which governs their decisions to follow or overtake. This herding behavior, along with intra-group cooperative-competitive interactions, forms the behavioral foundation of this study. Accordingly, the visual range of a two-wheeler is defined as a forward-facing sector area (Chen et al., 2024). The interacting agents within the forward visual sector are re-evaluated at each trajectory-propagation step according to their current positions and motion states. Therefore, the active pairwise interaction relationships may change over time and are not used to impose a fixed group membership. When other traffic participants fall within this sector, the time to the closest point of approach (TCPA) and the distance to the closest point of approach (DCPA) (Cheng et al., 2025) are computed to evaluate whether the interacting parties are converging and to infer their interaction intentions, and they are given by

$$\mathrm{TCPA} = t = -\frac{r \cdot v}{\|v\|^2} \quad (8)$$

$$\mathrm{DCPA} = \|r + v \cdot \mathrm{TCPA}\| \quad (9)$$

where r denotes the relative position vector between the two parties, and v represents the relative velocity vector.

The distance between the interacting participants at the closest-approach instant is calculated from their current relative motion states. TCPA and DCPA are used to identify whether the current relative motion corresponds to an imminent-convergence case. These thresholds are not used to determine the magnitude of the interaction force directly; instead, they provide a criterion for identifying close-interaction conditions and adjusting the direction of the interaction response. The magnitude of the interaction force varies continuously with the absolute value of TCPA through the exponential term in Eq.

Just Accepted

(10), so that the interaction increases as the closest-approach instant is approached and decays after that instant has passed. Furthermore, heterogeneous social-force parameters are introduced to characterize interactions with different participant types in mixed traffic.

$$F_{ij} = k_{b,c,p} \exp\left(-\frac{|\mathrm{TCPA}|}{\tau_{b,c,p}} \right) n \tag{10}$$

The absolute value of TCPA is used in the force-magnitude term to prevent the interaction force from increasing after the closest-approach instant, whereas the sign of TCPA is retained to distinguish approaching from post-closest-approach conditions.where n denotes a unit vector representing the direction of interaction. Its default orientation is defined from the centroid of the interacting object toward the centroid of the two-wheelers. When 0<TCPA≤2s and DCPA≤2m, the interacting pair is regarded as an imminent-convergence case, and the longitudinal component of the direction vector is reversed to represent a detouring response. Otherwise, the default interaction direction is retained. The parameters $k_{b,c,p}$ and $T_{b,c,p}$ correspond to the interaction intensity coefficients and reaction intensity coefficients when interacting with other two-wheelers, cars, and pedestrians, respectively.

3.3.3. Road Boundary Constraint Force

Considering the influence of road boundary constraints on two-wheelers,(11)defines the boundary repulsive force. A minimum distance threshold of 0.5 m between two-wheelers and road boundaries is imposed to ensure strict trajectory confinement. While traveling within lanes, lane-changing behavior introduces potential safety risks and psychological pressure on riders. Therefore, crossable lane line forces are additionally incorporated into the model to account for such interactions.

$$F_{b,l} = k_{b,l} \exp\left(\frac{d - d_l}{d} \right)^2 n_{\perp}, \quad d_l < d \tag{11}$$

where d_1 represents the actual minimum distance between the two-wheeler and the road boundary or lane line, d denotes the distance threshold below which road or lane forces begin to take effect, and $k_{b,l}$ indicates the repulsive strength coefficient corresponding to the road boundary and lane line constraints.

3.3.4. Decision Region Steering Force

Four rectangular turning decision zones are defined at intersections, activated when two-wheelers enter the intersection area. A steering force is applied according to the turning intention strength inferred from historical trajectories, as expressed in(12):

$$F_t = k_t (\theta_{\mathrm{desire}} - \theta) n_{\perp} \tag{12}$$

$$\theta_{\mathrm{desire}} = \arctan\left(\frac{v_{d,y}}{v_{d,x}} \right) \tag{13}$$

$$\theta = \arctan\left(\frac{v_y}{v_x} \right) \tag{14}$$

where k_t denotes the steering intensity coefficient that governs the application of steering force at intersections, θ_{desire} represents the desired heading angle, θ refers to the current actual heading angle, $v_{d,x}$ and $v_{d,y}$ indicate the x and y components of the desired velocity, respectively.

3.3.5. Kinematic Model of Two-Wheelers

The planar kinematic approximation is used to update the center position and heading of each two-wheeler in the RSFM branch. The planar motion state is represented by the position, speed, heading angle, and lateral acceleration. The yaw rate is determined by the lateral acceleration and current speed. The corresponding state-update relationships are expressed as follows:

$$\omega_t = \frac{a_{\mathrm{lat},t}}{v_t} \tag{15}$$

$$\theta_{t+1} = \theta_t + w_t \Delta t \tag{16}$$

$$x_{t+1} = x_t + v_t \cos(\theta_t) \Delta t \tag{17}$$

$$y_{t+1} = y_t + v_t \sin(\theta_t) \Delta t \tag{18}$$

where a_{lat} denotes the lateral acceleration, ω represents the yaw rate, θ denotes the heading angle, and v is the speed. The variables x and y represent the updated planar coordinates. Together, these variables define the planar state-transition process used by the RSFM to represent turning and lateral motion over the prediction horizon.

3.4. STTransformer Model Considering Group Temporal Dependencies

The input of the STTransformer consists of a 13-dimensional state feature vector that includes coordinate positions (x_i, y_i) ; velocities (v_x^i, v_y^i) and accelerations (a_x^i, a_y^i) in the x and y directions; longitudinal and lateral velocities $(v_{\mathrm{lon}}^i, v_{\mathrm{lat}}^i)$ and accelerations $(a_{\mathrm{lon}}^i, a_{\mathrm{lat}}^i)$;trigonometric components of the heading angle $(\sin\theta_i, \cos\theta_i)$; and the distance to road boundaries d_i , within the time range $i \in (1, T)$ (**Fig. 4**).

Capturing group herding behaviors and intra-group cooperative-competitive relationships is essential for accurately modeling and predicting two-wheeler trajectories. The scene-level group specifies the complete set of target two-wheelers to be predicted within the current interaction sample, whereas a target-specific local interaction graph is constructed separately for each predicted group member. For model computation, each two-wheeler in the scene-level group is successively treated as the prediction node. For a predicted two-wheeler i, surrounding traffic participants entering a 10 m neighborhood during the historical observation interval are selected as candidate interaction nodes. If more than 19 surrounding agents satisfy this condition, they are ranked according to their mean distance from the predicted two-wheeler, and the nearest 19 are retained. Together with the prediction node, each local interaction graph therefore contains at most 20 nodes. This value is a computational upper bound rather than a fixed two-wheeler group size. Accordingly, the scene-level group determines which two-wheelers require trajectory outputs, whereas the 10 m neighborhood determines which surrounding agents participate in the interaction modeling for a particular prediction target.

Just Accepted

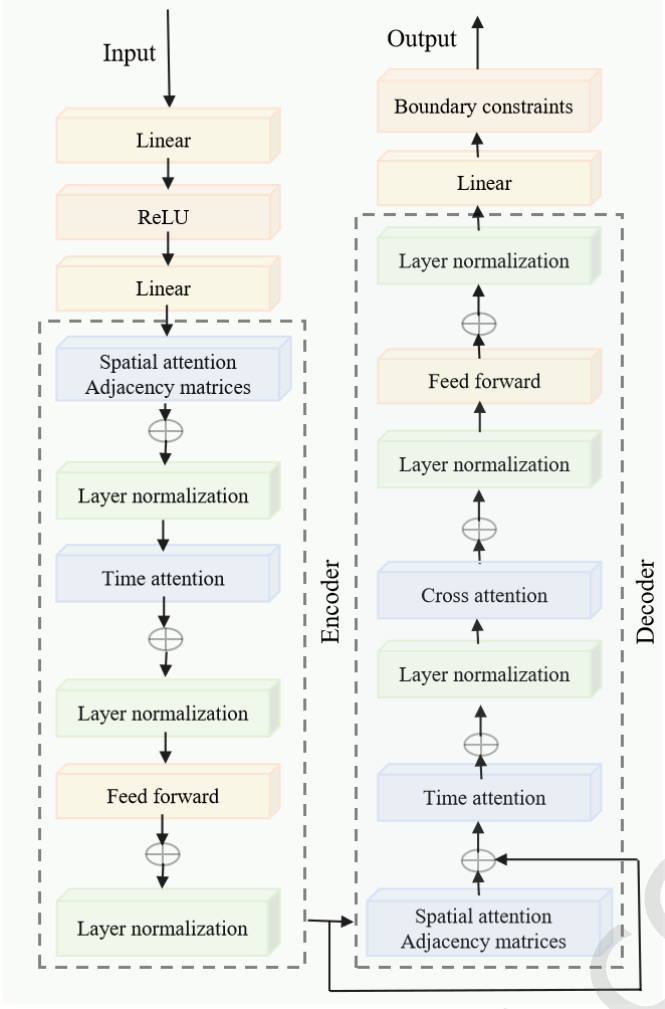


Fig. 4. Architecture of the STTransformer model.

Let N denote the actual number of nodes in one trajectory sample. At historical time step t, a Boolean adjacency matrix $A_t \in \{0,1\}^{N\times N}$ is constructed for the nodes present at that time step, with self-connections retained. If a node is absent, its corresponding adjacency entries remain zero. The sequence of adjacency matrices forms the three-dimensional spatiotemporal adjacency tensor $A_h \in \{0,1\}^{T_h\times N\times N}$. A visibility mask $M_h \in \{0,1\}^{T_h\times N}$ records the presence of each node over the historical observation interval. Consequently, the effective interaction topology can change as surrounding participants enter or leave the local neighborhood. Samples with different numbers of nodes are zero-padded only during batching, and padded nodes are excluded through zero-valued adjacency entries. The construction of the local spatiotemporal interaction graph is illustrated in **Fig.** 5.

A spatial self-attention mechanism is incorporated into the encoder to explicitly capture the spatial dependencies within two-wheeler groups. Additionally, an interactive feature fusion layer is developed to integrate intra-group dynamics with interactions involving surrounding heterogeneous traffic participants. Through the combination of a masked attention mechanism and a Multilayer Perceptron (MLP), this layer effectively strengthens the representation of group-level features.

$$f_{\text{inter}} = \text{MLP}\left(\text{concat}\left[f, \text{MaskedAtten}(f, f, f, A)\right]\right) \quad (19)$$

where Attention refers to the attention computation operation, and MLP denotes the multi-layer perceptron. This fusion mechanism

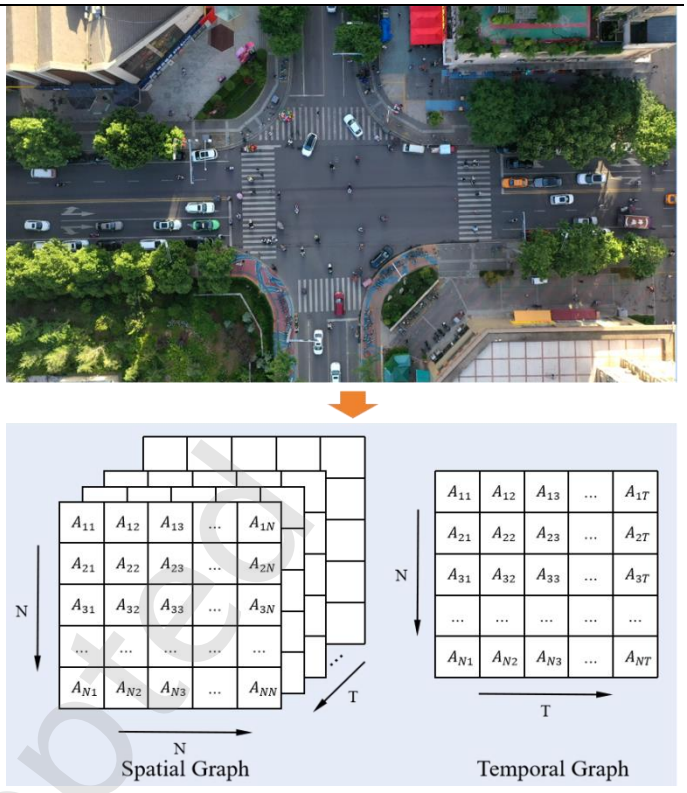


Fig. 5. Construction of the spatiotemporal graph.

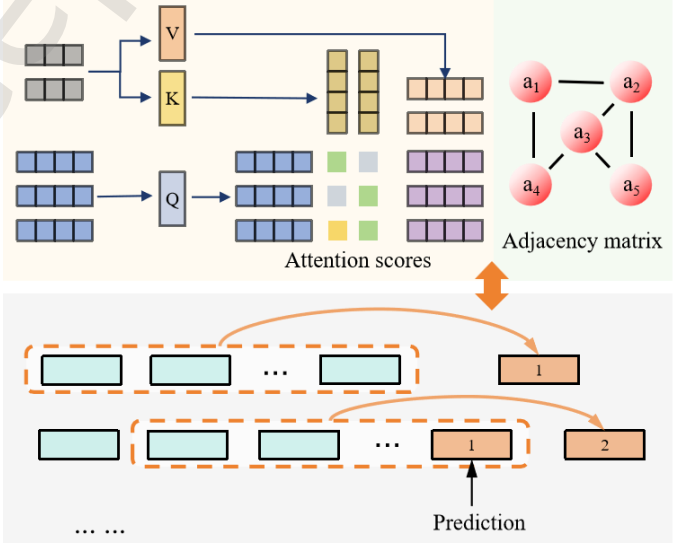


Fig. 6. Multi-step trajectory prediction for two-wheeler groups.

enables a more accurate representation of the interactive behaviors exhibited by the two-wheeler groups within mixed traffic flow.

The decoder adopts an architecture similar to that of the encoder and integrates cross-attention to fuse the memory representations M generated by the encoder, thereby enabling multi-step trajectory prediction. This module generates future trajectories sequentially, using the predicted results from the preceding time step as inputs for the next prediction(**Fig. 6**).

$$h_t^{\text{dec}} = \text{Decoder}\left(x_{t-1}^{\text{dec}}, M, A_t\right) \quad (20)$$

where x_{t-1}^{dec} represents the output from the previous decoding step, which corresponds to the ground-truth value during training and the model's own prediction during inference. Meanwhile, A_t denotes the adjacency matrix at time t ,capturing the dynamic spatial relationships between two-wheeler groups and surrounding traffic participants.

Just Accepted

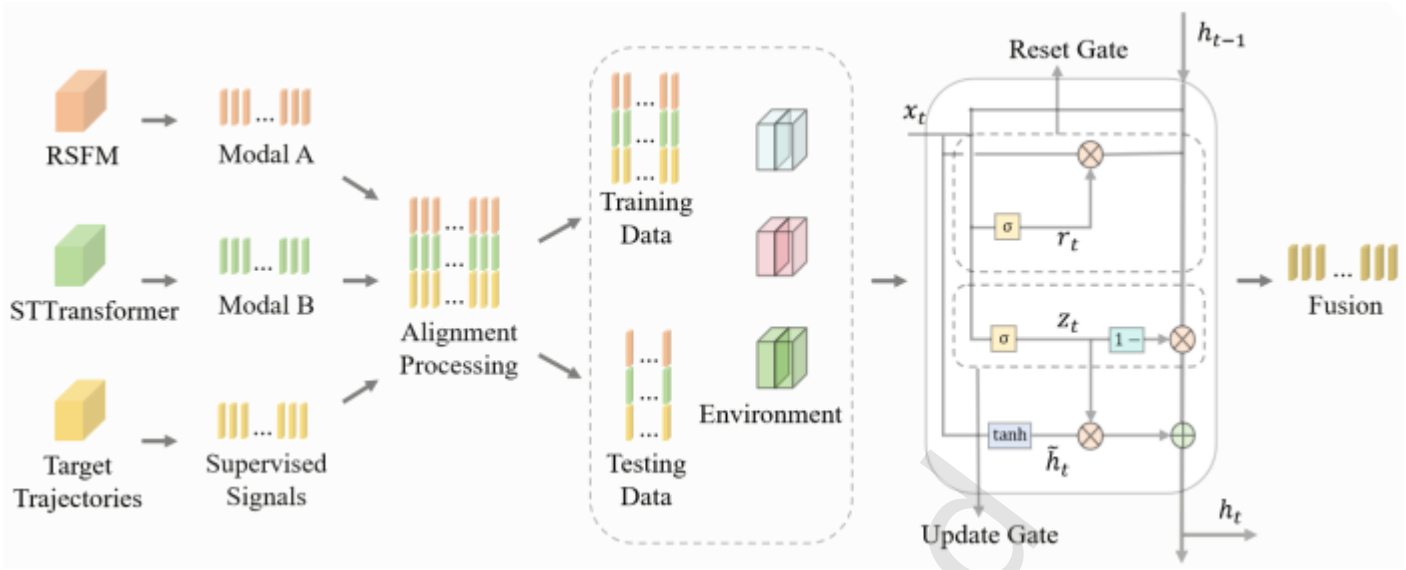


Fig. 7 Spatiotemporal information fusion process based on GRU.

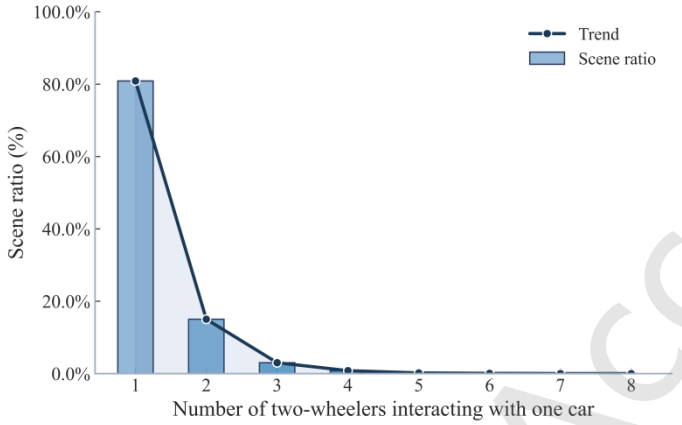


Fig. 8 Interaction scenarios between two-wheelers and vehicles.

The decoder initially transforms the input displacement vectors into high-dimensional feature representations and incorporates learnable positional encodings to capture temporal dependencies among two-wheeler groups. It then applies cross-attention to extract critical interactive features from the memory M ,providing contextual information for each decoding step and ensuring consistent and coherent trajectory evolution across the multi-step prediction process.

The distance from each observed trajectory point to the nearest road boundary is included in the input feature vector, allowing the encoder to incorporate road-boundary information during representation learning. During autoregressive inference, a geometricboundary-clearance correction is applied after each decoding step. If the distance between a predicted point and its nearest point on the boundary is smaller than the safety margin, the point is minimally displaced along the direction from the nearest boundary point toward the predicted point until the predefined safety margin is restored. Otherwise, the predicted point remains unchanged.

This deterministic operation uses only the known road geometry and is intended to maintain a predefined safety margin between predicted points and road boundaries during multi-step prediction. It does not constitute a learned kinematic constraint or guarantee complete dynamic feasibility. The model parameters are optimized using an (L2) loss applied over all predicted time steps.

3.5. Multi-Model Trajectory Prediction Fusion Method

The predicted trajectories generated independently by the RSFM and STTransformer are used as the two primary prediction inputs to the fusion model. Together with the scene complexity representation $E(\mathbf{C})$, these temporally aligned trajectory predictions form the input to a GRU-based multi-model fusion framework. Within this framework, the RSFM provides trajectory predictions incorporating interpretable spatial interactions and physical constraints, whereas the STTransformer provides trajectory predictions learned from historical spatiotemporal dependencies and structured interactions. A two-layer GRU recurrently integrates the complementary predictions of the two branches and the scene complexity representation to generate the final trajectory prediction for each target two-wheeler. (**Fig. 7**).

The GRU performs recurrent temporal fusion through its reset and update gates, which jointly regulate the recurrent hidden-state update based on the current fusion input and the previous hidden state. It should be noted that the reset and update gates operate on the entire joint fusion input, rather than assigning separate scalar weights to RSFM and STTransformer, respectively. In this study, adaptive fusion refers to the recurrent fusion state being conditioned on the current RSFM and STTransformer predictions, the scene complexity representation $E(\mathbf{C})$, and the previous hidden state. Accordingly, "adaptive" in this study does not denote explicit scene-dependent scalar weighting between the RSFM and STTransformer branches; rather, it denotes input-conditioned recurrent fusion in which the fusion state and output vary with the branch predictions, scene complexity representation, and recurrent history. Two gating units independently compute the reset signal r_t and update signal z_t based on current input x_t and previous hidden state h_{t-1} using corresponding weight matrices W_r and W_z .

The reset signal r_t determines which information from previous hidden states should be retained through selective filtering. Based on the current state information, the reset gate adaptively adjusts r_t to regulate the dependence on prior hidden states. The filtered historical information $r_t h_{t-1}$, together with the current input x_t , is used to compute the candidate hidden state h_t^c . The matrix W_h denotes the weight matrix corresponding to the candidate hidden state. For

Just Accepted

notational simplicity, products between a gate vector and a hidden-state vector denote element-wise multiplication throughout this subsection.

$$r_t = \sigma\left(W_r\left[h_{t-1}, x_t\right]\right) \tag{21}$$

$$h_t^c = \tanh\left(W_h\left[r_t h_{t-1}, x_t\right]\right) \tag{22}$$

The update signal z_t regulates the weighting between the historical state h_{t-1} and the candidate hidden state h_t^c during the final state update h_t. When z_t approaches 1, the network prioritizes the candidate state, which integrates the current fusion information from the RSFM and STTransformer. Conversely, when z_t approaches 0, the network preserves the continuity of the previous hidden state. Furthermore, the GRU fusion learning process employs an L2 loss function to optimize the model's predictive performance.

$$z_t = \sigma\left(W_z\left[h_{t-1}, x_t\right]\right) \tag{23}$$

$$h_t = \left(1 - z_t\right) h_{t-1} + z_t h_t^c \tag{24}$$

4 Case Study

4.1. Dataset

Table 1. Number of Trajectory Samples in the SIND Dataset

City	Pedestrian	Bicycle	Motorcycle	Tricycle	Car	Truck	Bus
Intersection1 (Tianjin)	1,652	2,807	3,360	479	4,792	107	51
Intersection2 (Changchun)	1,267	237	1,339	99	12,074	622	408
Intersection3 (Xi'an)	4,447	396	3,862	345	8,090	199	102

The proposed model is validated using the public SIND dataset, which includes diverse road users, such as cars, buses, trucks, and VRUs (Xu et al., 2022; Yang et al., 2025). Detailed information about the dataset is presented in **Table 1**.

Fig. 8 shows that multi-two-wheeler interaction scenarios constitute approximately 19.2% of all interaction cases in the Signalized INtersection Dataset (SIND). Most existing studies primarily address dyadic interactions between a vehicle and a single two-wheeler. However, predicting the trajectory of an individual two-wheeler alone cannot effectively capture the collective herding behaviors and internal interactive dynamics present within two-wheeler groups.

The STTransformer model was pre-trained using the Intersection1 subset of the SIND. The raw recordings were used as the splitting unit, and mutually exclusive sets of recordings were assigned to the training, validation, and test partitions before any agent-centered sliding-window samples were constructed. No raw recording was shared across these partitions. Sliding-window sampling was subsequently performed independently within each pre-assigned subset, and the generated samples were not randomly redistributed across subsets. Consequently, samples constructed with different two-wheelers as prediction nodes remain in the same subset as the raw recording from which they originate. In addition, sliding windows were restricted to the boundaries of their respective subsets; windows spanning two different subsets were not generated. Therefore, the complete historical observation and future prediction sequence of each sample originates from only one data subset.

Within the training and validation subsets, sliding windows with a step size of 5 were applied, generating 37,198 training samples and 5,196 validation samples. During testing, a step size of 20 was used to reduce temporal overlap and redundancy between adjacent test samples. The pre-trained STTransformer model was evaluated on the Intersection1 (6,472 samples), Intersection2 (1,080 samples), and Intersection3 (2,156 samples) datasets. The RSFM generated corresponding predictions on the same evaluation samples, and the trajectory outputs of the two branches were temporally and spatially aligned. These aligned branch predictions, rather than predictions generated from the STTransformer pre-training samples, were subsequently used together with the corresponding scene complexity representations to construct the input data for the GRU fusion stage. For the GRU fusion process, the resulting Intersection1, Intersection2, and Intersection3 fusion datasets were divided into training and testing sets with ratios of 7:3, 9:1, and 9:1, respectively. The average computation time of AdaFusionNet is 0.039 seconds, which is higher than that of RSFM (0.022 seconds) and STTransformer (0.031 seconds). Nevertheless, its prediction accuracy and average computation time indicate that it can support real-time trajectory-prediction applications under the evaluated hardware configuration. Our code is implemented using the PyTorch framework and was trained on an NVIDIA GeForce RTX 3090Ti.

4.2. Model Calibration

The RSFM parameters are calibrated using the PSO algorithm. The final optimized parameter values are presented in **Table 2**.

To improve the convergence efficiency of the STTransformer, a dynamic teacher forcing mechanism was implemented (Liu et al., 2025). The fusion model was trained using a two-layer GRU architecture. All training procedures employed the Adam optimizer with an initial learning rate of 10^{-4}.

4.3. Evaluation Metrics

The primary evaluation metrics include the root mean square error (RMSE), average displacement error (ADE), and final displacement error (FDE), which are given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{P}_i - P_i\right\|_2^2} \tag{25}$$

$$\mathrm{ADE} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{P}_i - P_i\right\|_2 \tag{26}$$

$$\mathrm{FDE} = \left\|\hat{P}_N - P_N\right\|_2 \tag{27}$$

Table 2. PSO-Based Parameter Optimization Results

Parameter	Value
τ_i	9.687
K_{bike}	0.112
T_{bike}	0.107
K_{ped}	0.103
T_{ped}	0.332
K_{car}	0.139
T_{car}	0.097
K_{boundary}	0.114
K_{lane}	0.102
k_t	9.869

where N denotes the total number of steps in the predictionsequence, $\hat{P}_i$ and P_i represent the predicted and true positions at step

Just Accepted

i ,respectively, while $\hat{P}_N$ and P_N indicate the positions at the final step of the predicted and true sequences, respectively.

4.4. Ablation Experiments

This study first employs the Intersection1 dataset to conduct ablation experiments on the RSFM and STTransformer under a 3-second trajectory prediction task. The RSFM ablation process is summarized in 错误!未找到引用源。. Interaction force $F_{interaction}$ with surrounding traffic participants, steering force F_{turn} in turning areas, and road network force F_{road} are gradually introduced. The prediction accuracy improves progressively with the inclusion of each social force component, achieving optimal performance when all forces are incorporated.

Table 3 Ablation Study of the RSFM

F_{drive}	$F_{interaction}$	F_{turn}	F_{road}	ADE ↓ / m	FDE ↓ / m	RMSE ↓ / m
√				1.0421	1.7652	1.1284
√	√			0.8603	1.6794	0.9034
√	√	√		0.7838	1.6372	0.8741
√	√	√	√	**0.7130**	**1.6085**	**0.8552**

The STTransformer ablation process is presented in 错误!未找到引用源。. The incorporation of spatiotemporal attention mechanisms leads to improved performance across all evaluation metrics compared to the baseline Transformer. When boundary constraints are additionally integrated, the model achieves optimal performance. These experimental results confirm the effectiveness of each functional module within the proposed model architecture.

Table 4 Ablation Study of the STTransformer Model

Transformer	Spatiotemporal Attention	Boundary	ADE↓/m	FDE↓/m	RMSE↓/m
√			0.9444	1.7405	1.2261
√	√		0.8447	1.6218	1.0493
√	√	√	**0.7319**	**1.4948**	**0.8561**

The two ablation experiments further provide complementary evidence for the contribution of interaction-aware group modeling. In the RSFM branch, introducing the dynamic interaction force improves prediction by explicitly incorporating interactions with surrounding traffic participants. In the STTransformer branch, the improvement obtained after introducing spatiotemporal attention indicates that structured spatial and temporal relations among the predicted two-wheelers and surrounding agents provide additional predictive information beyond motion-history modeling alone.

To further evaluate the fusion strategy and the contribution of the scene complexity representation, five variants were compared using the same RSFM and STTransformer trajectory predictions and the same data partition. Simple averaging directly averages the two branch predictions without learned fusion parameters. Fixed weighting combines the two predictions using a constant fusion weight selected on the validation set. MLP fusion incorporates thesame scene complexity representation $E(\mathbf{C})$ but replaces the recurrent fusion module with a feed-forward MLP. AdaFusionNet-V0 retains the two-layer GRU fusion architecture but removes $E(\mathbf{C})$, whereas the complete AdaFusionNet incorporates both recurrent temporal fusion and the scene complexity representation. Detailed results are reported in **Table** .

AdaFusionNet achieves the lowest ADE, FDE, and RMSE at both prediction horizons. Compared with AdaFusionNet-V0, the complete model reduces all six reported errors, with improvements ranging from approximately 9.4% to 17.2%. Its advantage over MLP fusion further demonstrates the contribution of recurrent hidden-state modeling, while the comparisons with simple averaging and fixed weighting show that learned nonlinear fusion is more effective than static combination strategies. These results indicate that $E(\mathbf{C})$ provides complementary information about local interaction complexity for the fusion process rather than directly acting as a predefined gating rule. These results demonstrate the benefit of conditioning the recurrent fusion process on local scene information.

Table 3 Ablation of the fusion strategy and scene complexity representation on Intersection 1.

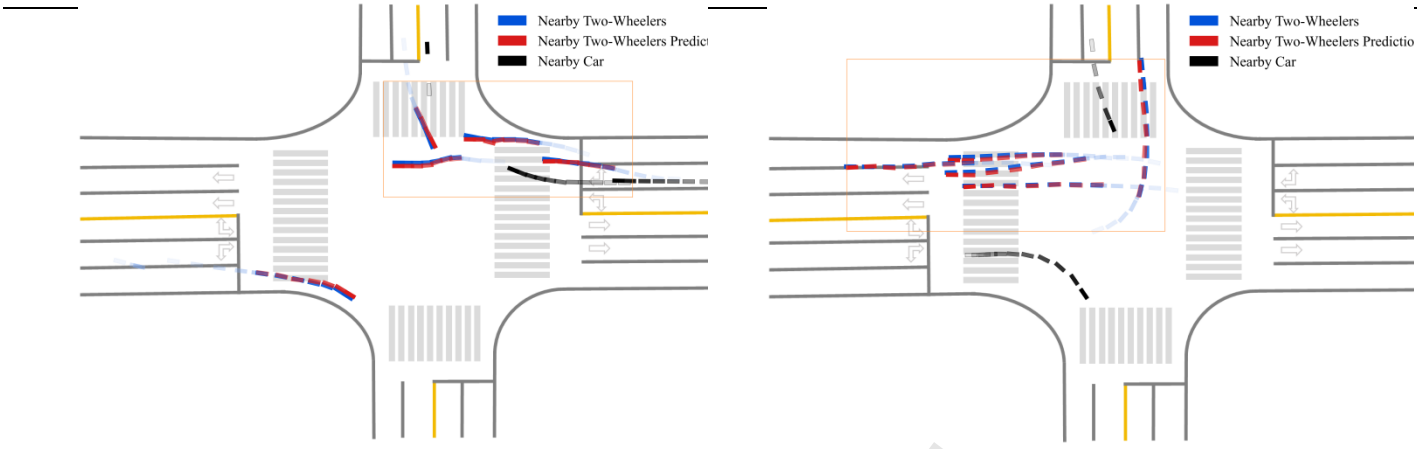
Method	Temporal fusion	Complexity representation	2s (20 frames) ↓			3s (30 frames) ↓		
			ADE / m	FDE / m	RMSE / m	ADE / m	FDE / m	RMSE / m
Simple averaging			0.4286	0.8227	0.4879	0.6638	1.4146	0.8017
Fixed weighting			0.4169	0.7885	0.4723	0.6375	1.3518	0.7662
MLP fusion		√	0.3978	0.7564	0.4591	0.6108	1.2987	0.7445
AdaFusionNet-V0	√		0.4059	0.7412	0.4964	0.5903	1.2504	0.7311
AdaFusionNet	√	√	0.3677	0.6398	0.4109	0.5334	1.0711	0.6189

Note: "Scene complexity representation" denotes $E(\mathbf{C})$, which consists of the numbers and mean speeds of surrounding pedestrians and vehicles. "√" indicates that the corresponding component is included.

Table 4 Comparison of evaluation metrics (m) over different prediction horizons at Intersection1.

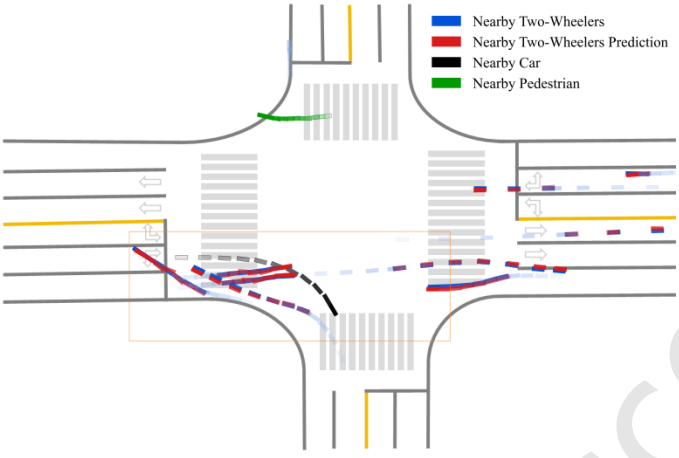
	SIND/Intersection1	2s (20 frames) ↓			3s (30 frames) ↓		
		ADE / m	FDE / m	RMSE / m	ADE / m	FDE / m	RMSE / m
Comparison	Vanilla-LSTM	0.8357	1.9831	1.1486	1.5288	3.7214	2.1358
	Social-LSTM	0.7856	1.8583	1.0737	1.4312	3.4689	1.9882
	GRIP++	0.7397	1.4777	0.9276	1.1355	2.3479	1.4416
	Social GAN	0.3855	0.8913	0.5840	0.6879	1.6461	1.0764
	FJMP	0.3694	0.8514	0.5523	0.6017	1.5218	0.9384
Ablation & Comparison	RSFM	0.4391	0.8759	0.5045	0.7130	1.6085	0.8552
	STTransformer	0.4808	0.9322	0.5539	0.7319	1.4948	0.8561
	AdaFusionNet	0.3677	0.6398	0.4109	0.5334	1.0711	0.6189

Just Accepted

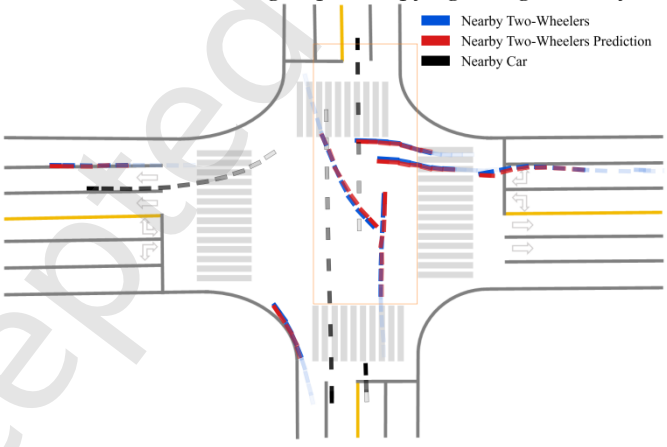


(a) Two-wheeler group cut off by the car

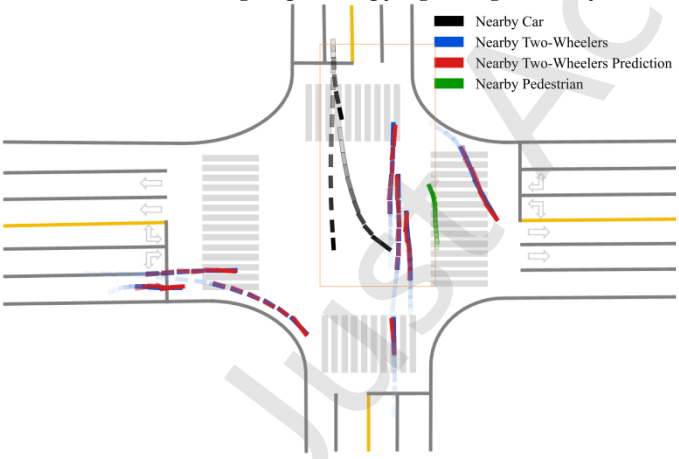
(b) Two-wheeler group 1 occupying the right-of-way



(c) Two-wheeler group 2 occupying the right-of-way



(d) Two-wheeler group 1 yielding the right-of-way



(e) Two-wheeler group 2 yielding the right-of-way



(f) Two-wheeler group moving against traffic flow

Just Accepted

Fig. 9 Comparison of trajectory prediction results in interaction scenarios.

They should not, however, be interpreted as evidence of explicit branch-specific weight changes between the RSFM and STTransformer, because the GRU gates operate on the complete fusion input rather than producing separate scalar weights for the two prediction branches.

4.5. Comparative Experiments

4.5.1 *Quantitative Results*

The proposed AdaFusionNet is evaluated against several established baseline models for trajectory prediction tasks:

① Vanilla-LSTM: A separate LSTM is trained for each agent, where each network processes only the individual's past trajectory to learn motion patterns and predict future paths.

② Social-LSTM (Alahi et al., 2016): Information is shared between multiple LSTMs through pooling layers to capture interactions within the local neighborhood.

③ GRIP++ (Li et al., 2019): Interactions between nearby agents are represented using graphs, with multiple graph convolutional blocks employed for feature extraction, followed by an LSTM for prediction.

④ Social GAN (Gupta et al., 2018): Trajectories are predicted through an adversarial training framework consisting of a discriminator and a generator, with diversity loss incorporated to

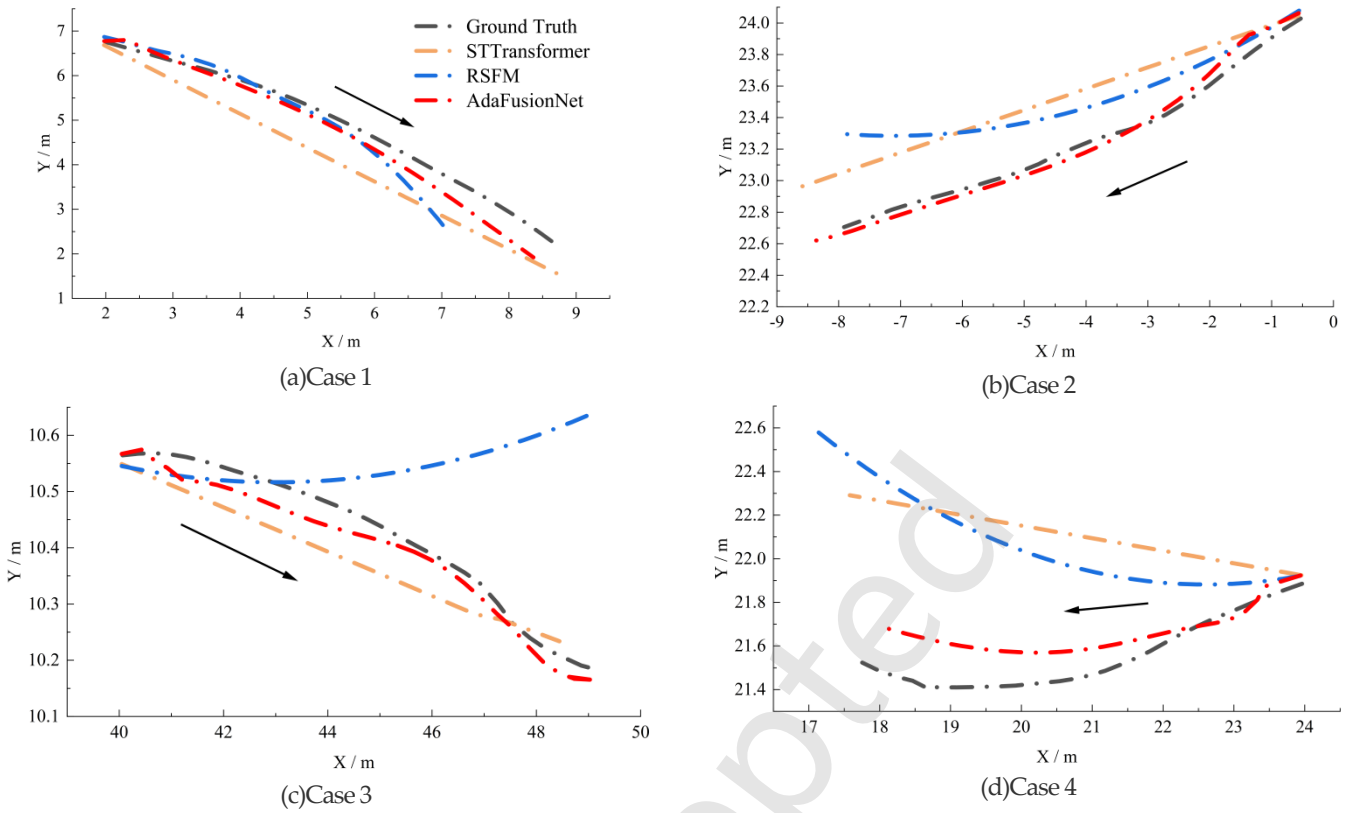


Fig. 10 Comparison of trajectory prediction results in non-group scenarios.

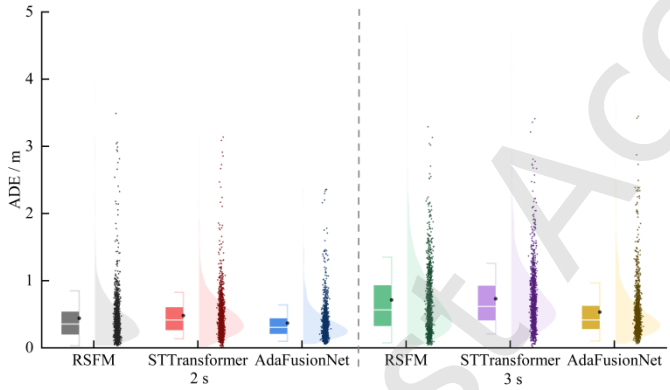


Fig. 11. Distribution of trajectory prediction errors.

encourage a range of plausible trajectory predictions.

⑤ FJMP (Rowe et al., 2023): FJMP models multi-agent interactions as a sparse directed graph, prunes the graph into a DAG, decomposes the joint prediction task accordingly, and decodes trajectories using a DAGNN. FJMP generates six candidate trajectory modes, and its performance is evaluated using minADE and minFDE over the six predicted modes.

Table reports the ADE, FDE, and RMSE results of eight methods for two-wheeler trajectory prediction over 2-second and 3-second horizons. The experimental results show that AdaFusionNet consistently achieves the best performance among all comparative models. For the 2-second prediction, compared with traditional temporal models (Vanilla-LSTM, Social-LSTM, and GRIP++), the proposed method reduces all three metrics by an average of more than 50%, demonstrating its effectiveness in motion sequence modeling. When compared with the GAN-based Social GAN, AdaFusionNet achieves a 4.6% improvement in ADE and notable reductions of 28.2% and 29.6% in FDE and RMSE, respectively. Compared with FJMP, AdaFusionNet obtains a marginally lower ADE (0.3677 m versus 0.3694 m) and achieves reductions of 24.9% and 25.6% in FDE and RMSE, respectively, indicating lower endpoint and overall displacement errors. Relative to the individual RSFM and STTransformer models, all evaluation metrics improve by more than 16%, validating the benefits of integrating social interaction modeling and spatiotemporal feature learning. For the 3-second prediction, although the errors of all comparison models increase, AdaFusionNet maintains the lowest error values. The improvements over Social GAN, FJMP, RSFM, and STTransformer exceed 11.4%, with RMSE reduced by 34.0% compared with FJMP, demonstrating the advantage of the proposed model under the longer prediction horizon.

Overall, the comparative advantage of AdaFusionNet at the 2-second horizon is metric-dependent, whereas its performance advantage becomes more evident at the 3-second horizon, with improvements exceeding 11.4% over the compared methods. This performance advantage primarily arises from the complementary multi-model trajectory-prediction and recurrent fusion framework. On one hand, the model effectively captures microscopic motion correlations by representing group interaction effects among two-wheelers. On the other hand, it incorporates the scene complexity representation $\boldsymbol{E(C)}$, providing complementary local interaction information for the fusion process. Furthermore, the approach integrates the behavioral modeling strength of the RSFM with the spatiotemporal dependency learning capability of the STTransformer, thereby establishing a more comprehensive representational framework. Notably, the larger accuracy gain observed in the 3-second prediction highlights the advantage of the proposed framework for longer-horizon trajectory prediction.

4.5.2. Case Analysis

Two-wheelers in road environments often demonstrate lower compliance with traffic regulations, as well as distinct herding

Just Accepted

behaviors and intra-group cooperative-competitive dynamics. Therefore, this study first evaluates the model's performance in scenarios characterized by complex group interactions. To provide future motion information for autonomous vehicles interacting with groups of two-wheelers, simultaneous trajectory predictions are performed for all two-wheelers surrounding the AV. Using the AdaFusionNet model, the trajectories of two-wheeler groups at the intersection are predicted for the subsequent 3 s.

In the complex task of predicting the trajectories of two-wheeler groups interacting with the subject vehicle, the AdaFusionNet model exhibits strong predictive capability. When the subject vehicle attempts to find a gap and pass through a dense flow of two-wheelers, the model accurately predicts the overall evolution trend of the surrounding traffic flow(**Fig.** 9a).(**Fig.** 9b-**Fig.** 9c)depict a typical competitive leader-follower interaction, where the model forecasts that the two-wheeler group, while avoiding the subject vehicle, follows the leading rider and accelerates to cross the potential conflict zone ahead of the vehicle. Conversely, (**Fig.** 9d-**Fig.** 9e) illustrate a cooperative leader–follower dynamic, where the model predicts that the two-wheeler group follows the leading rider and performs coordinated avoidance, yielding to the subject vehicle. Moreover, the model demonstrates the ability to predict unconventional or even non-compliant behaviors, successfully identifying scenarios in which the two-wheeler group collectively moves against the traffic flow under the influence of the leader (**Fig.** 9f). These results demonstrate that AdaFusionNet effectively captures diverse micro-level interaction patterns, including competitive, cooperative, and non-compliant behaviors, while accurately predicting the motion trajectories of two-wheeler groups in highly dynamic environments. This validates the importance of incorporating group interaction modeling in mixed traffic scenarios to enhance the accuracy and reliability of trajectory prediction.

Second, to examine the prediction performance of the proposed model under relatively simple motion conditions, a detailed comparative analysis was conducted under non-group-interactive scenarios. Distinct model architectures exhibit notable differences in predictive performance when handling such scenarios **(Fig.** 10**a-Fig.** 10**d).**

The STTransformer model demonstrates a pronounced tendency to produce overly smooth and nearly linear trajectory predictions. This behavior may be associated with its emphasis on global spatiotemporal dependencies and the use of an MSE objective, which tends to penalize large pointwise deviations and may favor conservative predictions in ambiguous motion conditions.

In contrast, the kinematics-based RSFM is more responsive to local nonlinear motion changes and can closely follow short-term trajectory variations. This advantage arises from its explicit modeling of instantaneous motion states and interaction-related physical factors. However, because the RSFM primarily relies on local motion information and does not explicitly model long-term temporal dependencies, small deviations generated during successive prediction steps may gradually accumulate. As the prediction horizon increases, such accumulated deviations can lead to directional drift and larger trajectory errors, thereby limiting its performance in longer-horizon prediction.

The effectiveness of AdaFusionNet stems from the recurrent integration of complementary trajectory predictions rather than a simple static combination of its constituent models. The recurrent fusion mechanism integrates the trajectory predictions generated by the STTransformer and RSFM together with the scene complexity representation $E(\mathbf{C})$, allowing the final prediction to incorporate complementary information from both branches over the prediction sequence. Combined with the results in interactive scenarios, these observations demonstrate that AdaFusionNet maintains competitive prediction performance across both complex interaction scenarios and relatively simple motion conditions.

As illustrated by the statistical visualization results in **Fig. 11**, AdaFusionNet exhibits improved prediction accuracy compared with the individual prediction branches. This enhancement is visually evident from the downward shift of both the black dot (mean) and the white line (median) in its boxplot. Relative to the other models, the violin plot of AdaFusionNet shows a shorter, narrower, and more compact upper portion of the error distribution, with fewer dispersed

sample points. These distributional characteristics are consistent with the benefits of the proposed multi-model fusion mechanism. By incorporating the scene complexity representation $E(\mathbf{C})$, the fusion model receives complementary information about the local traffic context in addition to the trajectory predictions generated by the two branches. Together with recurrent temporal fusion, this information contributes to the lower prediction errors observed for AdaFusionNet. AdaFusionNet has a lower mean and median error than the individual prediction branches. Its error distribution also exhibits a more compact upper portion, indicating fewer large displacement errors within the evaluated test samples.

4.6. Cross-Validation in Multi-Intersection

To further examine the prediction performance of AdaFusionNet across the evaluated intersections, additional within-intersection evaluations were conducted using the Intersection 2 and Intersection 3 subsets of the SIND dataset. The corresponding results are reported in **Table** 5 and **Table** 6.

At Intersection2, under the 2-second prediction horizon, AdaFusionNet achieved substantial performance gains over most baseline models. Compared with sequence-based models, such as Vanilla-LSTM and Social-LSTM, ADE and FDE improved by more than 50%, while improvements over the graph-based GRIP++ model exceeded 60%. Although FJMP achieved a 7.6% lower ADE, AdaFusionNet reduced FDE and RMSE by 20.2% and 26.2%, respectively. This result indicates that FJMP has an advantage in short-horizon average displacement, whereas AdaFusionNet produces lower endpoint and overall displacement errors. Under the 3-second prediction horizon, the model further exhibited comprehensive superiority, achieving more than 11% improvement across all evaluation metrics compared with all baseline models, including Social GAN, FJMP, RSFM and STTransformer.

A similar performance pattern was observed at Intersection3. Under the 2 s horizon, FJMP achieved a slightly lower ADE than AdaFusionNet, with values of 0.9446 m and 0.9640 m, respectively. In contrast, AdaFusionNet achieved lower FDE and RMSE, with reductions of 15.4% and 33.8% relative to FJMP. At the 3 s horizon, AdaFusionNet obtained the lowest values for all three metrics among the compared methods. When the prediction horizon was extended to

Just Accepted

Table 5 Comparison of evaluation metrics (m) over different prediction horizons at Intersection2

	SIND/Intersection2	2s (20 frames) ↓			3s (30 frames) ↓		
		ADE / m	FDE / m	RMSE / m	ADE / m	FDE / m	RMSE / m
Comparison	Vanilla-LSTM	1.5225	3.5230	1.9131	2.7036	6.4252	3.4705
	Social-LSTM	1.4521	3.3761	1.8249	2.5651	6.0395	3.2778
	GRIP++	1.8926	3.7108	2.5673	2.2826	4.6845	3.0940
	Social GAN	0.6682	1.4675	1.1282	1.1488	2.6639	1.9270
	FJMP	**0.6621**	1.4029	1.0603	1.0181	2.5573	1.8351
Ablation & Comparison	RSFM	0.7675	1.4330	0.8524	1.1946	2.5974	1.4006
	STTransformer	0.8538	1.6774	0.9888	1.2959	2.6102	1.5071
	AdaFusionNet	0.7162	**1.1198**	**0.7824**	**0.9048**	**1.4717**	**1.0134**

Table 6 Comparison of evaluation metrics (m) over different prediction horizons at Intersection3

	SIND/Intersection3	2s (20 frames) ↓			3s (30 frames) ↓		
		ADE / m	FDE / m	RMSE / m	ADE / m	FDE / m	RMSE / m
Comparison	Vanilla-LSTM	1.7822	3.9686	2.2265	3.1501	7.2342	3.9471
	Social-LSTM	1.6729	3.7533	2.0985	2.9804	6.8871	3.7544
	GRIP++	2.2822	4.5051	3.2217	2.9129	5.9973	3.9623
	Social GAN	0.9621	2.0960	1.6465	1.6228	3.6982	2.7089
	FJMP	**0.9446**	1.9703	1.6121	1.5920	3.4124	2.5062
Ablation & Comparison	RSFM	1.5227	3.0787	1.7521	2.4302	5.2904	2.8746
	STTransformer	1.5131	3.0940	1.7668	2.4078	5.1639	2.8514
	AdaFusionNet	0.9640	**1.6669**	**1.0677**	**1.4089**	**2.8812**	**1.6356**

Table 7 Leave-one-intersection-out evaluation on unseen SIND intersections.

Training intersections	Unseen test intersection	Model	2s (20 frames) ↓			3s (30 frames) ↓		
			ADE / m	FDE / m	RMSE / m	ADE / m	FDE / m	RMSE / m
I2+I3	I1	RSFM	0.4852	0.9915	0.5489	0.8185	1.8900	0.9587
		STTransformer	0.5707	1.1457	0.6492	0.9119	1.9298	1.0427
		FJMP	0.4270	1.0242	0.6451	0.7347	1.9388	1.1223
		AdaFusionNet	**0.4115**	**0.7339**	**0.4553**	**0.6209**	**1.2832**	**0.7086**
I1+I3	I2	RSFM	0.8888	1.6136	1.0007	1.4084	2.9922	1.6667
		STTransformer	1.0664	2.1521	1.2133	1.6536	3.4402	1.8899
		FJMP	**0.7892**	1.7368	1.2575	1.2665	3.2733	2.2333
		AdaFusionNet	0.8136	**1.3124**	**0.8794**	**1.0740**	**1.8028**	**1.1837**
I1+I2	I3	RSFM	1.7709	3.6606	2.0097	2.9527	6.6024	3.4380
		STTransformer	1.9035	4.0717	2.1802	3.2048	7.1365	3.7211
		FJMP	**1.1071**	2.4905	1.9313	2.0473	4.5760	3.1503
		AdaFusionNet	1.1182	**2.0103**	**1.2150**	**1.7386**	**3.6620**	**1.9447**

3 s, the model's advantage became more pronounced, achieving performance gains ranging between 11.5% and 55.3% relative to all baseline models, further demonstrating its prediction advantage at the longer prediction horizon across the evaluated intersections.

These results show that AdaFusionNet consistently achieves lower FDE and RMSE across the evaluated intersections, although FJMP obtains slightly lower 2 s ADE at Intersections 2 and 3. The advantage of AdaFusionNet becomes more evident at the 3 s horizon, where it achieves the lowest ADE, FDE, and RMSE among the compared methods.

To further evaluate cross-intersection transfer under an unseen-intersection setting, a strict three-fold leave-one-intersection-out evaluation was conducted. In each fold, data from two intersections were used for model development, while the remaining intersection was excluded from training and used exclusively as the unseen test intersection. The three folds were I2+I3→I1, I1+I3→I2, and I1+I2→I3. The complete quantitative results are reported in **Table** 7.

AdaFusionNet achieves the best performance in 16 of the 18 reported metric–horizon combinations. It obtains the lowest FDE and RMSE at both the 2 s and 3 s horizons for all three held-out intersections and achieves the lowest ADE, FDE, and RMSE at the 3 s horizon in all three folds. For held-out Intersections 2 and 3, FJMP obtains slightly lower 2 s ADE values than AdaFusionNet, whereas AdaFusionNet achieves lower FDE and RMSE.

These results provide additional evidence of cross-intersection transferability across the three evaluated SIND intersections.

5 Conclusion

In this study, the AdaFusionNet model is proposed, which employs a GRU-based fusion framework integrating RSFM to capture real-time spatial interactions and STTransformer to model long-term temporal dependencies. This integration enhances the trajectory prediction accuracy of two-wheeler groups in complex traffic environments. Numerical experiments conducted using the SIND

Just Accepted

signalized intersection data demonstrate that the proposed approach achieves improved prediction accuracy compared with the selected comparison methods.

1)AdaFusionNet demonstrates competitive performance across both 2-second and 3-second prediction horizons, with consistent advantages in FDE and RMSE across the evaluated intersections. Notably, its performance advantage becomes more evident as the prediction horizon extends: while the comparative advantage at the 2-second horizon depends on the evaluation metric, the accuracy improvement exceeds 11% under the more challenging 3-second condition. This improvement verifies the effectiveness of the proposed framework for longer-horizon trajectory prediction.

2)AdaFusionNet accurately captures the behavioral dynamics of two-wheeler groups in trajectory prediction. The model effectively predicts both complex coordinated and individual behaviors within groups while also maintaining competitive prediction performance in sparse scenarios involving simple linear motions of individual two-wheelers. This adaptive capability enables the model to handle varying degrees of group interaction intensity and heterogeneous motion patterns, while the leave-one-intersection-out results further demonstrate transfer performance across the three evaluated SIND intersections.

3) AdaFusionNet exhibits a more concentrated prediction-error distribution, with fewer large displacement errors in the evaluated samples. This characteristic allows the model to provide more stable trajectory estimates across different interaction scenarios, thereby providing useful motion information for downstream autonomous-driving decision-making.

Several research directions warrant further exploration. Future work can be extended to incorporate additional environmental factors, such as weather conditions and road surface characteristics, which can substantially impact two-wheeler behavior and trajectory prediction accuracy. Future work will also investigate the integration of AdaFusionNet with autonomous-vehicle planning and control modules for closed-loop evaluation.

Author contributions

Lan Yang: Conceptualization, Methodology, Funding acquisition, Supervision, Writing – review & editing.

Yuanhao Liu: Methodology, Validation, Formal analysis, Project administration, Writing – review & editing, Correspondence.

Guangyue Qu: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Visualization, Writing – original draft.

Yueyang Wang: Methodology, Validation, Writing – review & editing.

Shan Fang: Supervision, Methodology, Writing – review & editing.

Shoucai Jing: Supervision, Resources, Writing – review & editing.

Xiangmo Zhao: Writing – review & editing.

Replication and data sharing

The trajectory data and accompanying code associated with this study are publicly available in the AdaFusionNet replication package through the ETS-Data repository (Dataset ID: ETS-Data-26-00098; DOI: https://doi.org/10.26599/ETSD.2026.9190113).

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 52472446 and 52602727) and the Natural Science Foundation of Shaanxi Province (Grant No. 2026JC-YXQN168).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

Alahi, A., Goel, K., Ramanathan, V., et al., 2016. Social LSTM: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 961–971.

Chen, H., Liu, Y., Hu, C., Zhang, X., 2023. Vulnerable road user trajectory prediction for autonomous driving using a data-driven integrated approach. IEEE Transactions on Intelligent Transportation Systems, 24(7), 7306–7317.

Chen, H., Liu, Y., Zhao, B., et al., 2022. Vision-based real-time online vulnerable traffic participants trajectory prediction for autonomous vehicle. IEEE Transactions on Intelligent Vehicles, 8(3), 2110–2122.

Chen, K., Zhao, X., Huang, Y., et al., 2025. SocialTrans: Transformer-based social intentions interaction for pedestrian trajectory prediction. Physica A: Statistical Mechanics and its Applications, 130435.

Chen, X., Li, S., Yang, W., et al., 2024. Enhanced microsimulation framework for right-turning vehicle-pedestrian interactions at signalized intersection. Simulation Modelling Practice and Theory, 134, 102930.

Cheng, K., Li, W., Zhou, Y., et al., 2025. A novel fuzzy comprehensive evaluation method for ship collision risk. Ocean Engineering, 332, 121462.

Feng, S., Liu, F., Zhou, S., et al., 2025. Dynamic obstacle avoidance strategy for autonomous vehicles based on LSTM model trajectory prediction under urban roads. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 09544070241299283.

Fu, L., Zhang, Y., Qin, H., et al., 2023. A modified social force model for studying nonlinear dynamics of pedestrian-e-bike mixed flow at a signalized crosswalk. Chaos, Solitons & Fractals, 174, 113813.

Gao, H., Su, H., Cai, Y., et al., 2021. Trajectory prediction of cyclist based on dynamic Bayesian network and long short-term memory model at unsignalized intersections. Science China Information Sciences, 64, 172207.

Gupta, A., Johnson, J., Li, F.-F., et al., 2018. Social GAN: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2255–2264.

Huang, L., Wu, J., You, F., et al., 2016. Cyclist social force model at unsignalized intersections with heterogeneous traffic. IEEE Transactions on Industrial Informatics, 13(2), 782–792.

Just Accepted

Huang, Z., Wang, J., Pi, L., et al., 2021. LSTM-based trajectory prediction model for cyclist utilizing multiple interactions with environment. Pattern Recognition, 112, 107800.

Li, J., Ni, Y., Sun, J., 2023a. A two-layer integrated model for cyclist trajectory prediction considering multiple interactions with the environment. Transportation Research Part C: Emerging Technologies, 155, 104304.

Li, J., Ni, Y., Sun, J., 2024a. Considering cyclists' aggressiveness and bounded rationality: A self-learning motion prediction model for cyclists. IEEE Transactions on Intelligent Vehicles.

Li, L., Wang, X., Yang, D., et al., 2023b. Real-time heterogeneous road-agents trajectory prediction using hierarchical convolutional networks and multi-task learning. IEEE Transactions on Intelligent Vehicles, 9(1), 4055–4069.

Li, M., Chen, T., Du, H., 2022. Trajectory prediction of cyclist based on spatial-temporal multi-graph network in crowded scenarios. Electronics Letters, 58(3), 97–99.

Li, T., Jiao, W., 2024. Geometric algebra based Kalman filter trajectory prediction. In: Proceedings of the 3rd International Symposium on Aerospace Engineering Systems (ISAES), 103–108.

Li, X., Yang, L., Li, R., et al., 2026. EE-RL: Vision language guided reinforcement learning with explorer and expert model for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 32082–32092.

Li, X., Ying, X., Chuah, M.C., 2019. GRIP++: Enhanced graph-based interaction-aware trajectory prediction for autonomous driving. arXiv preprint arXiv:1907.07792. https://doi.org/10.48550/arXiv.1907.07792

Li, Z., Chen, Z., Li, Y., Xu, C., 2024. Context-aware trajectory prediction for autonomous driving in heterogeneous environments. Computer-Aided Civil and Infrastructure Engineering, 39(1), 120–135.

Liu, Q., Shen, X., Liu, C., et al., 2025. Enhancing the generalization capability of 2D array pointer networks through multiple teacher-forcing knowledge distillation. Journal of Automation and Intelligence, 4(1), 29–38.

Liu, Y.-C., Jafari, A., Shim, J.K., et al., 2022. Dynamic modeling and simulation of electric scooter interactions with a pedestrian crowd using a social force model. IEEE Transactions on Intelligent Transportation Systems, 23(9), 16448–16461.

Liu, Z., Zhong, N., Chen, J., et al., 2024. A modeling method for two-dimensional two-wheeler driving behavior during severe conflict interaction at intersections. Accident Analysis & Prevention, 205, 107668.

Rowe, L., Ethier, M., Dykhne, E.-H., Czarnecki, K., 2023. 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 (CVPR), 13745–13755.

Rudenko, A., Palmieri, L., Lilienthal, A.J., et al., 2018. Human motion prediction under social grouping constraints. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 3358–3364.

Wang, W., Zhou, H., Lo, J.T.Y., et al., 2024. A modified social force model for pedestrian-bicycle mixed flows and its application on evaluating the conflict risk in shared roads. Physica A: Statistical Mechanics and its Applications, 643, 129788.

Wang, Y., Liu, Z., Zuo, Z., et al., 2019. Trajectory planning and safety assessment of autonomous vehicles based on motion prediction and model predictive control. IEEE Transactions on Vehicular Technology, 68(9), 8546–8556.

Wei, C., Wu, G., Barth, M.J., et al., 2024. 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 (CVPR), 7115–7124.

Xu, D., Shang, X., Peng, H., Li, H., 2023. MVHGN: Multi-view adaptive hierarchical spatial graph convolution network based trajectory prediction for heterogeneous traffic-agents. IEEE Transactions on Intelligent Transportation Systems, 24(6), 6217–6226.

Xu, R., Li, J., Zhang, S., et al., 2025. Interactive trajectory prediction for autonomous driving based on Transformer. Mechanical Sciences, 16, 87–97.

Xu, Y., Shao, W., Li, J., et al., 2022. SIND: A drone dataset at signalized intersection in China. In: Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), 2471–2478.

Yang, L., Liu, S., Feng, S., et al., 2025. Generation of critical pedestrian scenarios for autonomous vehicle testing. Accident Analysis & Prevention, 214, 107962.

Zernetsch, S., Trupp, O., Kress, V., et al., 2021. Cyclist trajectory forecasts by incorporation of multi-view video information. In: Proceedings of the 2021 IEEE International Smart Cities Conference (ISC2), 1–7.

Zhang, R., Ma, X., Zhang, C., et al., 2024. GA-FCFNN: A new forecasting method combining feature selection methods and feedforward neural networks using genetic algorithms. Information Sciences, 669, 120566.

Zhou, H., Ren, D., Xia, H., et al., 2021. AST-GNN: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction. Neurocomputing, 445, 298–308.

Zhu, J., Chen, S., Tu, W., et al., 2019. Tracking and simulating pedestrian movements at intersections using unmanned aerial vehicles. Remote Sensing, 11(8), 925.

Just Accepted

Author biography

	Lan Yang received the Ph.D. degree from Chang'an University, Xi'an, China. She is currently a Professor and Ph.D. Supervisor with the School of Information Engineering, Chang'an University. She also serves as Deputy Chief Engineer at the Shaanxi Engineering Technology Research Center for Intelligent Road Traffic Detection and Equipment. Her research interests include autonomous driving, artificial intelligence, and intelligent transportation systems.
	Yuanhao Liu received the B.S. degree in transportation from the School of Aviation, Inner Mongolia University of Technology, in 2025. He is currently pursuing the M.S. degree at the School of Information Engineering of Chang'an University, Xi'an, China. His research interests include autonomous vehicle testing and evaluation.
	Guangyue Qu received the M.S. degree from the School of Information Engineering, Chang'an University, Xi'an, China. His research interests include autonomous driving testing and evaluation, deep learning, and intelligent transportation systems.
	Yueyang Wang received the B.Eng. degree in Transportation from Shanghai University of Engineering Science, Shanghai, China, and the M.Sc. degree in Automotive Engineering from the University of Leeds, Leeds, U.K. He is currently pursuing the Ph.D. degree with the Institute for Transport Studies, University of Leeds. His research interests include road user interaction modelling, pedestrian–vehicle interaction, autonomous driving, and human behaviour modelling.
	Shan Fang received the B.S. degree in Computer Science from Jilin University, Changchun, China, in 2016, and the Ph.D. degree in Transportation Information Engineering and Control from Chang'an University, Xi'an, China, in 2024. He is currently a Lecturer with the School of Information Engineering, Chang'an University. His research interests include artificial intelligence methods, autonomous driving, and smart infrastructure monitoring and maintenance.
	Shoucai Jing received the Ph.D. degree and is currently an Associate Professor and M.S. Supervisor with the School of Information Engineering, Chang'an University, Xi'an, China. His research interests include cooperative control of intelligent connected vehicles, active vehicle safety, autonomous driving, intelligent transportation systems, and autonomous vehicle testing and evaluation.

Just Accepted



Xiangmo Zhao received the Ph.D. degree from Chang'an University, Xi'an, China. He is currently a Professor and Ph.D. Supervisor with the School of Information Engineering, Chang'an University. He is the leader of the National Key Discipline of Traffic Information Engineering and Control and a recipient of the State Council Special Allowance. His research interests include intelligent connected vehicles, autonomous driving, vehicle–road cooperative systems, intelligent transportation systems, and autonomous vehicle testing and evaluation.

Just Accepted