Drones and Autonomous Vehicles

SCIEPublish

Article

MUGI-Net: A Group-Aware Pedestrian Trajectory Prediction Model for Autonomous Vehicles from First-Person View

Rongrong Ni, Sijie Yang, Menyun Du and Biao Yang *

Wang Zheng Institute of Micro-Electronics, Changzhou University, Changzhou 213000, China; nirongrong@cczu.edu.cn (R.N.); s24060809002@smail.cczu.edu.cn (S.Y.); S23060809038@smail.cczu.edu.cn (M.D.)

* Corresponding author. E-mail: yb6864171@cczu.edu.cn (B.Y.)

Received: 12 March 2026; Revised: 26 March 2026; Accepted: 10 April 2026; Available online: 22 April 2026

ABSTRACT: With the rapid development of autonomous driving, first-person view (FPV) pedestrian trajectory prediction has emerged as a key research direction to improve transportation system safety and operational efficiency. However, current studies ignore inter-pedestrian group information and long- and short-term dependence, leading to error accumulation at medium and long temporal horizons. To address these problems, we propose an FPV pedestrian trajectory prediction model dubbed MUGI-Net (Mixture of Universals and Group Interaction Network). It adopts a group pooling mechanism to adaptively aggregate group nodes and build sparse intra- and inter-group interaction graphs to fuse group interaction information. Afterward, it employs a Mixture of Universals (MoU) structure that combines MoF (Mixture of Feature Extractors) and MoA (Mixture of Architectures) to capture short-term dynamics and long-term dependencies simultaneously. Extensive experiments on the JAAD and PIE datasets show that MUGI-Net reduces the 1.5 s prediction MSE by 5% compared with the state-of-the-art AANet, and achieves the best performance on multiple key metrics, which is beneficial for autonomous driving in mixed traffic scenarios.

Keywords: First-person view; Trajectory prediction; Group interaction; Hybrid temporal encoding

1. Introduction

With the ongoing evolution of autonomous driving technology and intelligent transportation systems (ITS), pedestrian safety has become an important research topic in traffic safety [1]. Especially against the backdrop of the accelerated implementation of advanced driver assistance and autonomous driving functions, how to achieve reliable prediction of pedestrians' future motion states, enabling vehicles to plan online and adjust motion strategies promptly to reduce potential collision risks, has become a key issue that needs to be broken through [2]. Direct value for improving road traffic safety is not only provided by pedestrian trajectory prediction, but also by fundamental support for refined, intelligent regulation of smart city traffic management and traffic flow [3,4]. In this context, pedestrian trajectory prediction from the first-person view (FPV) has gradually attracted attention: this paradigm models directly based on on-board forward perception information, avoiding to a certain extent the geometric errors and cumulative deviations introduced by view transformation and coordinate mapping in bird's-eye view (BEV) methods [5–8].

However, the FPV scenario is characterized by more significant dynamic changes in imaging perspective, occlusion, and scale transformation, making it more difficult to stably characterize the motion-interaction relationships between pedestrians in the scene [9,10]. To improve FPV prediction performance, existing studies have proposed multi-modal prediction frameworks that usually combine historical trajectories and relevant contextual information to generate multiple feasible future trajectories [11–13]. Dynamic changes in the FPV scenario are addressed by these methods to some extent. However, they remain insufficient for modeling dynamic interactions at the group level. Individual historical motion and local spatial cues are focused on by most works, making it difficult to fully characterize the time-evolving interaction effects among pedestrians, thereby limiting the generalization ability and the medium- to long-term prediction accuracy in complex traffic environments [14,15].

Overall, existing FPV pedestrian trajectory prediction methods still face the following challenges: (i) The interaction relationships among pedestrians are highly nonlinear and time-varying. Existing models usually focus on a single trajectory of the target pedestrian, lack spatiotemporal feature fusion, and make it difficult to quantify the contribution of individuals to group behavior, thereby limiting explicit modeling capability for complex group interactions and the accurate characterization of social relationships and group interaction laws. (ii) Existing models mostly rely on relatively static pattern assumptions, and traditional temporal modeling structures tend to smooth out segment differences and lose short-term details, making it difficult to depict the rapid switching of pedestrians' local behavior patterns, resulting in continuous error accumulation in long temporal prediction and the inability to balance the capture of local dynamic patterns and the global temporal dependence.

To address the above problems, this paper proposes a method called Mixture of Universals and Group Interaction Network (MUGI-Net), as shown in Figure 1. The main innovations are summarized as follows:

(1) Aiming at the problem that traditional methods usually only model around the target pedestrian's trajectory and are difficult to depict the influence of group interaction, this paper proposes a group pooling mechanism. It realizes dynamic group division by fusing information on relative distance, velocity, and motion direction among pedestrians. It constructs intra-group and inter-group interaction graphs, explicitly modeling pedestrian cooperative behavior at the group level, thereby improving interaction expression and effectively reducing the complexity of graph structure modeling.
(2) Aiming at the problem that traditional temporal models are difficult to balance short-term dynamic capture and long-range dependence modeling, this paper proposes a Mixture of Universals (MoU) structure. It realizes adaptive short-term feature extraction via the Mixture of Feature Extractors (MoF). It fuses Mamba, convolution, and self-attention mechanisms via the Mixture of Architectures (MoA), thereby achieving multi-scale temporal dependence modeling.
(3) To improve the multi-modal trajectory prediction capability, this paper introduces an Iterative Refinement Module (IRM) based on intention anchors in the decoding stage. By treating learnable intention anchors as query vectors and combining a multi-layer Transformer architecture to perform layer-by-layer interactions and semantic updates on the historical trajectory context, it gradually aligns with the global motion pattern prior to the sample's conditional modality, thereby generating more stable and diverse future trajectory prediction results.

Experimental results show that the method in this paper achieves excellent performance in the FPV pedestrian trajectory prediction task: on the JAAD dataset, the model achieves MSE_15 = 153 in the 1.5 s prediction task, which is about 5% lower than the current state-of-the-art AANet method (161); on the PIE dataset, the model achieves MSE_10 = 33 in the 1.0 s prediction task, and outperforms most existing methods in indicators such as MSE_15, CMSE and CFMSE, verifying the effectiveness of the proposed group interaction modeling and hybrid temporal encoding structure. The rest of this paper is organized as follows: Section 2 introduces related research; Section 3 elaborates on the proposed pedestrian trajectory prediction model and its key modules; Section 4 presents the experimental settings and analyzes the results; and Section 5 summarizes the paper and outlines future research directions.

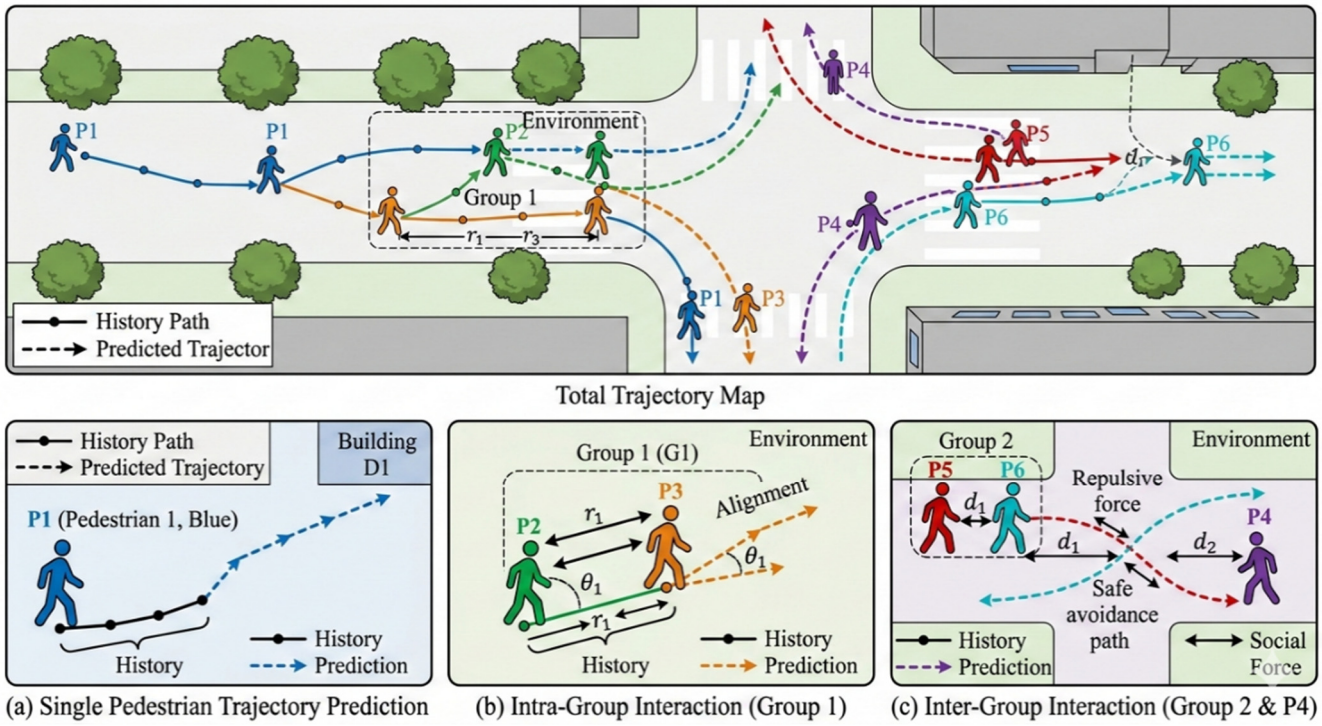

Figure 1. Schematic diagram of pedestrian trajectory prediction. (**a**) Single pedestrian, (**b**) intra-group pedestrian interaction, (**c**) inter-group pedestrian interaction.

2. Related Work

This section systematically reviews existing research on pedestrian trajectory prediction and the temporal modeling techniques used to model it. Mainstream methods and the existing limitations of related tasks are outlined.

2.1. Pedestrian Trajectory Prediction

In recent years, deep learning has advanced significantly in FPV trajectory prediction. For instance, many studies combine CNNs with RNNs to predict future positions using historical trajectory data, capturing spatiotemporal relationships [16–20]. Due to the randomness of pedestrian trajectories, Xu et al. [21] proposed a multi-modal framework combining LSTM and CNN to generate diverse trajectories, applicable to FPV scenarios. As intelligent transportation development advances, FPV prediction also requires contextual information. Niu et al. [22] proposed the AANet model, which processes pedestrian trajectories and ego-vehicle motion using a two-stream architecture to enhance scene understanding by learning about agent interactions.

Modeling pedestrian social and group interactions is another key direction. SHENet correlates individual and group trajectories with environmental information, while BiTraP captures bidirectional group motion correlation [17]. Recent studies integrated collision awareness into the graph Transformer [23], but lack FPV-specific optimizations. Dynamic subclass-balancing contrastive learning addresses long-tail trajectory issues [24]. Federated learning combined with dual-channel destination guidance protects privacy and improves trajectory rationality [25]. The TPPO predictor with a pseudo-Oracle module provides prior information for multi-modal prediction [26]. Despite covering multiple dimensions, previous studies did not sufficiently explore group interaction, and pedestrian analysis was isolated. To address this, pedestrians are grouped by relative direction and distance, and intra- and inter-group interactions are explored to improve FPV prediction performance.

2.2. Temporal Encoding

In recent years, the temporal encoding module has become a core component in trajectory prediction models. Early works are mainly based on RNNs (e.g., LSTM, GRU), which encode observed trajectory sequences into hidden states to model long-term dependence between historical and future motions [27,28]. With Transformers, work increasingly uses self-attention to replace/supplement RNN encoding for unified temporal modeling of long-range dependence and multi-agent interaction [29,30]. Yuan et al.'s AgentFormer [27] applies self-attention in temporal and agent dimensions to encode the "socio-temporal" structure and improve prediction performance. To enhance scene adaptability, an adaptive progressive Transformer adjusts attention weights based on scene features [31]. For fast cross-scene adaptation, meta-inverse reinforcement learning [32] and online multi-source transfer learning [33,34] enhance generalization from different perspectives by mining motion reward functions and extracting universal temporal features, respectively.

Traditional temporal encoding has limitations: RNNs are limited in modeling long-term dependencies, Transformers weaken local details, and linear state-space models lack expressiveness. To address this, this paper proposes a MoU-structured temporal encoding method that balances local representation and global temporal dependence while maintaining controllable complexity.

3. Methodology

The overall framework of the proposed FPV pedestrian trajectory prediction model is detailed in this section. To address problems of unstable grouping, insufficient temporal feature modeling, and single-modal prediction bias in first-person view scenarios, the model comprises four core components: a group interaction module, a MoU-structured temporal module, and a decoder with intention anchor refinement. Specifically, the group interaction module accurately groups pedestrians and models fine-grained intra- and inter-group interactions in FPV scenes, the temporal module efficiently captures multi-scale temporal dependencies via a hybrid structural design, and the decoder generates diverse and reasonable future trajectories through iterative intention refinement. The design details of each module are introduced sequentially below. The symbols used in this work are shown in Table 1.

Table 1. Symbols used in this work.

Symbol	Definition
T_h/T_f	Historical/Future observation time length
X_i/V_i	Position/Velocity information of the *i*-th pedestrian
G_{ped}	Pedestrian graph
V_{ped}	Pedestrian node set
E_{ped}	Edge set of social interactions
$\hat{Y}$	Predicted trajectory of the model

3.1. Overall Model Architecture

Given a historical horizon of T_h, the goal is to predict K possible trajectories of pedestrians in the future T_f time steps, denoted as $Y_j = \{y_{T_h+1}, \ y_{T_h+2}, \ ...y_{T_h+T_f}\}, j \in K$. As shown in Figure 2, the encoder module processes historical trajectories and inter-group information via positional encoding and a temporal module to produce representations suitable for model computation. In the temporal module, features are extracted using token embeddings and Mamba layers. Then, in the decoder module, convolutional layers and attention mechanisms are used to further capture local features and critical information. Self-attention and cross-attention mechanisms help the model focus on different parts of the input and on relationships

between modalities, and the final trajectory prediction results are generated by a feed-forward network. The key modules are elaborated below.

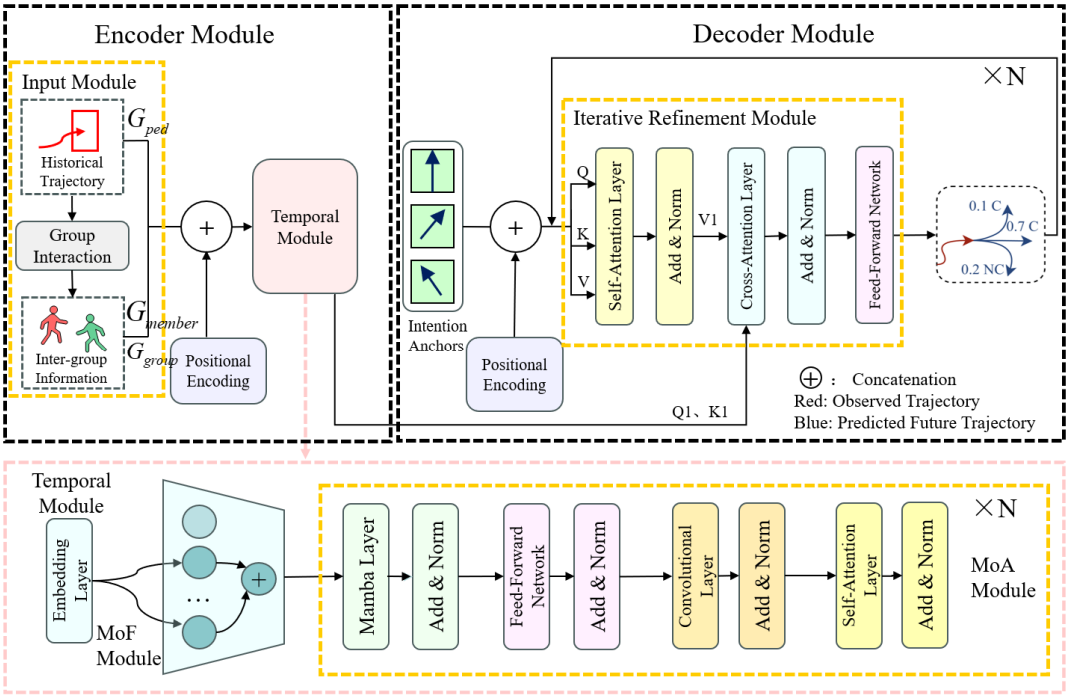

Figure 2. Pipeline of the proposed method.

3.2. Group Interaction Module

FPV pedestrian trajectory prediction is affected by both pedestrian motion and ego-motion. Thus, traditional group interaction modeling for BEV may lead to unstable grouping, identity confusion, and distorted interactions. To address the above challenges, we propose a group interaction module, as shown in Figure 3. Its overall process is elaborated as follows: (1) the historical trajectories are used for dynamic grouping; (2) the grouping results are used to construct intra-group interaction graphs and inter-group interaction graphs; (3) intra-group interaction features and inter-group interaction features are extracted from the two types of graphs respectively; (4) these two types of interaction features are then fused with historical trajectory features; (5) the fused features are used as the input of the temporal encoding module and enter MoF and MoA; (6) the encoding results are then sent to the decoder to complete the prediction.

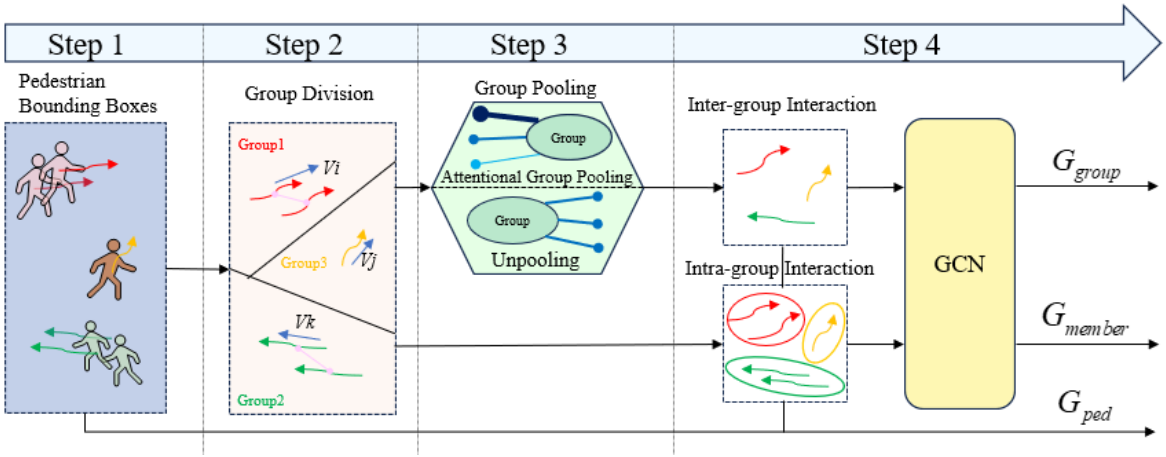

Figure 3. Flowchart of the group interaction module.

The above processes allow the group interaction module to effectively extract group-level social interaction features while preserving individual identity information, providing richer, structured input for the subsequent temporal encoding module and improving the overall performance of FPV pedestrian trajectory prediction. As shown in Step 1 of Figure 3, the set of bounding boxes of all pedestrians at time t within the observation time range T_h is expressed as:

$$X = \{x_1, x_2, \cdots, x_{T_h}\} \in R^{T_h \times N_x} \tag{1}$$

where x_t denotes the bounding box at the time step t, and N_x is its dimension.

3.2.1. Group Division

In the pedestrian trajectory prediction task, each node must retain its original identity information and describe the dynamic attributes of group behavior in the scene, thereby providing a basis for restoring the original identity information in subsequent steps. Existing studies have failed to adequately address pedestrian behavioral characteristics, especially motion direction.

Group division based solely on Euclidean distance calculations is prone to trajectory crossing or identity confusion [30]. To solve this problem, A method that fuses Euclidean distance and velocity direction is proposed, accounting for the influence of both motion velocity and spatial distance. As shown in Step 2 of Figure 3, because the distance, velocity difference, and direction angle between individuals in the same group may have significant characterizing effects, feature similarity among all pedestrians is calculated using historical trajectories, Euclidean distances, and velocity magnitudes and directions for each pedestrian. Based on the obtained similarity measurement, pedestrian combinations that may belong to the same group can be effectively identified. For this purpose, A pairwise distance matrix D is defined, and a set of group membership indicator variables P is introduced, which are specifically expressed as follows:

$$D_{ij} = \| F_\varphi(X_i, V_i) - F_\varphi(X_j, V_j) \|, \forall i, j \in \{1, 2, \cdots, N\} \tag{2}$$

$$P = \{pair(i,j) | i, j \in [1, 2, \cdots, N], i \neq j, D_{ij} \leq \Pi\} \tag{3}$$

where F_Φ is a learnable convolutional layer for learning deep group features. X_i and V_i represent position information and velocity, respectively. Π is a learnable threshold parameter. $Pair(\cdot,\cdot)$ represents pedestrians that may belong to the same group.

Subsequently, group indicators are constructed by analyzing the relationships among group members. In this context, the variable k represents the index of the k-th group, which is the union of each paired set (i, j). The key point is that there is no overlap between members across groups. Therefore, the group index G is defined as a set containing all members of a specific group:

$$G = \{G_k | G_k = \bigcup_{(i,j) \in P} \{i, j\}, G_a \cap G_b = \emptyset \text{ for } a \neq b\} \tag{4}$$

3.2.2. Group Pooling

To deeply explore the interaction relationships among pedestrians, A graph network is adopted for modeling [31] in this study. In this model, it is necessary to maintain the identity index information of each graph node to ensure that relevant nodes are included and irrelevant nodes are excluded. However, existing methods often use an average aggregation across the entire group, which is inconsistent with the study's goal. From the first-person view, average aggregation will further amplify the perspective scale difference: the bounding boxes of pedestrians closer to the camera change more drastically, and undifferentiated averaging is likely to lead to the group representation being dominated by "large-scale changes in the near field".

To achieve this goal, a new method called attentional group pooling and unpooling is proposed. As shown in Step 3 of Figure 3, first, pedestrian nodes are grouped based on common behavior patterns and surrounding information sets. During this grouping process, the features of the corresponding nodes are aggregated into a single representative node for each pedestrian group, guided by attention weights. The aggregated group features are stacked for use in subsequent modules to learn sparse connections in the spatiotemporal graph. Attentional pooling is used to select the most representative features for each pedestrian node. The influence of group members is determined by their velocity, with higher velocity leading to greater influence and being assigned larger weights. Using the above method, we construct a graph with significantly fewer nodes than the input graph. The clustered trajectory features Z are defined as follows:

$$Z = \{Z_k | k \in [1, 2, \cdots, K]\}, Z_k = W_k \sum_{i \in G_k} X_i \tag{5}$$

where K is the total number of groups, W is the attention weight, and X_i is the feature of each member.

To restore the grouped graph structure to the original scale, an unpooling operation is required. This method can predict each agent's trajectory based on the agent's output features and fusion information. Since the convolution operation on zero-vector nodes cannot effectively capture group attributes [35,36], this study evenly distributes group features Z to N relevant group-member nodes and reintroduces geometric information related to camera motion in the unpooling stage, so that each member node obtains the same group behavior information. The pedestrian group unpooling operator is defined as follows:

$$\bar{X} = \left\{ \bar{X}_n | n \in [1, 2, \cdots N] \right\} \tag{6}$$

$$\bar{X}_n = Z_k / N \quad where \quad n \in G_k \tag{7}$$

3.2.3. Intra- and Inter-Group Interaction

Social interaction is integrated into the sparsely connected spatiotemporal graph network at the group level. As shown in Step 4 of Figure 3, by providing three different graph-structure data types (intra-group interaction, inter-group interaction, and historical trajectory) to the same base model, the model can extract diverse features. The hierarchical structure for mapping intra-group interaction features is at the individual level. The nodes remain pedestrians. The edges only connect members within the same group. The main modeling focuses on the local coordination relationships among group members, including speed consistency, formation maintenance, mutual following, and collision-avoidance constraints. The hierarchical level for constructing interaction features between groups is the group level; the nodes have become "group nodes". The edges represent the interactions between different groups. The main modeling focuses on the high-level relationships between groups, such as yielding, overlapping, competing for space, and overall avoidance. The pedestrian graph $G_{ped} = (V_{ped}, E_{ped})$ is defined as a set $V_{ped} = \{X_n \mid n \in [1, 2, \ldots, N]\}$ of pedestrian nodes and a set $E_{ped} = \{e\_{ij} \mid i, j \in [1, 2, \ldots, N]\}$ of edges representing pairwise social interactions. The intra-group interaction graph is defined as $G_{member} = (V_{ped}, E_{member})$, which consists of a set V_{ped} of pedestrian nodes and a set E_{member} of edges representing pairwise social interactions among group members, where $E_{member} = \{e_{ij} \mid i, j \in [1, 2, \ldots, N], (i, j) \subset G_k, k \in [1, 2, \ldots, N]\}$. This graph enables pedestrian nodes to learn avoidance norms within the group while maintaining their formation and motion direction. From the first-person view, intra-group interaction also needs to maintain invariance to "overall formation translation/scaling caused by camera following": by introducing an ego-motion correction term in the intra-group edge features, the model can focus more on the changes in relative motion (relative velocity/relative direction) among members rather than global view drift.

Inter-group interaction is also crucial for learning social norms between different groups. The inter-group interaction graph is defined as $G_{group} = (V_{group}, E_{group})$, where nodes $V_{group} = \{X_k \mid k \in [1, 2, \ldots, K]\}$ represent the features of each group and edges $E_{group} = \{e_{p,q} \mid p,q \in [1, 2, \ldots, K]\}$ represent the interactions between paired groups. In the first-person scenario, inter-group relationships are affected by the field of view and perspective depth: distant groups appear denser and smaller in the image, while near groups are sparser and exhibit greater scale changes. For this reason, inter-group edge features can encode both "relative orientation/scale change trend" and "ego-motion consistency", thereby improving the ability to model interactions across distance scales.

In this framework, weights are shared by the model, thus reducing the number of parameters and improving the overall performance. Subsequently, the model's output features are aggregated into agents, and the group fusion module is used to predict the probability distribution of future trajectories. The predicted trajectory $\hat{Y}$ generated by the group fusion module F_ψ is expressed as:

$$\hat{Y} = F_\Psi\Big(\underbrace{F_\theta(X, G_{ped})}_{Historical\ Trajectory}, \underbrace{F_\theta(X, G_{member})}_{Intra-group\ Interaction}, \underbrace{F_\theta(\bar{X}, G_{group})}_{Inter-group\ Interaction}\Big) \tag{8}$$

where F_ψ and F_θ are learnable parameters, which are randomly initialized at the beginning.

3.3. Temporal Module

The temporal module is the core component of the Mixture of Universals (MoU) model, by which short- and long-term dependence relationships in temporal data are effectively captured. As shown in Figure 4, unlike the traditional Transformer and Mamba temporal encoding methods, this model combines two key components: Mixture of Feature Extractors (MoF) and Mixture of Architectures (MoA), which are used to extract short-term features and model long-term dependencies, respectively.

First, in the MoF module (Section 3.3.1), the model divides the trajectory sequence into multiple time patches using a sliding window and employs multiple sub-extractors to adaptively extract features across different patches, thereby capturing dynamic change features over a short time range more accurately. Then, in the MoA module (Section 3.3.2), the model performs hierarchical modeling on the extracted temporal features by fusing multiple structures such as Mamba layers, convolutional layers, self-attention layers, and feed-forward networks, to learn both local dependence relationships and global long-term dependence simultaneously, thus improving the model's expressive ability for complex trajectory patterns. Through the above design, the temporal module can effectively model the multi-scale temporal dependence of pedestrian trajectories while maintaining computational efficiency and providing richer, more stable temporal feature representations for subsequent trajectory prediction.

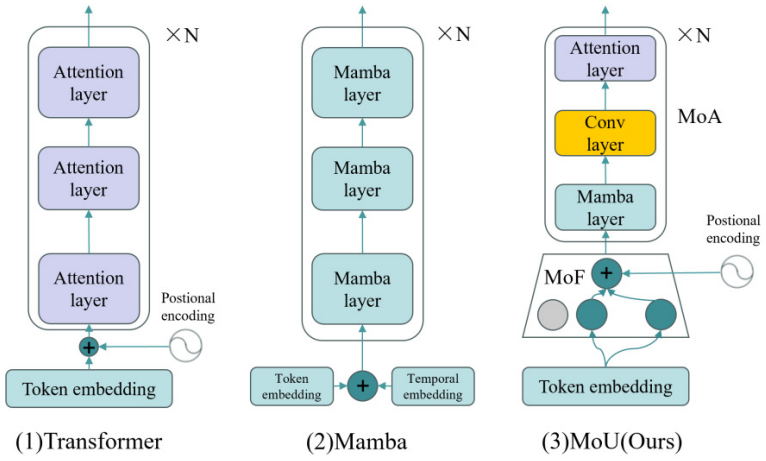

Figure 4. Comparison of MoU with traditional temporal encoding modules.

3.3.1. Mixture of Feature Extractors (MoF)

Traditional time-series patch-embedding methods (such as linear projection) use the same transformation parameters for all patches, ignoring differences in feature affinity across patches due to semantic context. This homogenization process will lead to the loss of part of the context information and limit the accurate representation of short-term details.

Features from each time-series patch are extracted by MoF, which is an adaptive feature extractor. In MoF, the input time series is divided into multiple patches through a sliding window. For each patch, MoF is processed by a set of Sub-Extractors to produce the patch's feature representation. The operation of MoF can be expressed as follows:

$$X_{rep} = MoF(X_p) = \sum_{i=1}^{n} R_i(X_p) F_i(X_p) \tag{9}$$

where X_p is the input time series patch. $F_i(X_p)$ is the i-th sub-extractor, which generates the patch's feature representation. $R_i(X_p)$ is the routing function, which is used to selectively activate the most suitable sub-extractor to ensure computational efficiency and adaptability.

3.3.2. Mixture of Architectures (MoA)

Although Transformers can model global long-term dependencies, they are computationally intensive; Mamba is computationally efficient, but there is a risk of information loss in long-term prediction. Dependence relationships are modeled step by step by MoA, from local to global, through a hierarchical hybrid architecture that balances efficiency and performance.

MoA is used to capture long-term dependence relationships in time series. As shown in Figure 5, the MoA structure includes four levels: Mamba layer, Feed Forward layer, Convolution layer, and Self-Attention layer, each of which captures different aspects of long-term dependence. The calculation of MoA is as follows:

(1) Mamba layer: First, it selectively processes time-varying dependence relationships as follows:

$$x' = \sigma(Conv1D(Linear(x))) \tag{10}$$

$$z = \sigma(Linear(x)) \tag{11}$$

where σ represents the activation function (usually SiLU), $Conv1D$ represents the one-dimensional convolution operation, and Linear represents the linear transformation.

(2) Self-Attention layer: Then, MoA captures global long-term dependence through the self-attention layer:

$$x_{att} = FeedForward(Attention(Q,K,V)) \tag{12}$$

where the calculation of the self-attention layer is:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{13}$$

where Q is the query vector, K is the key vector, V is the value vector, and d_k is the dimension of the key.

(3) Convolution layer: It is used to expand the receptive field, so that the dependence relationships between different patches can be learned more comprehensively:

$$x_{conv} = Conv(x_{ffn}; k, s, p, c_{out}) \tag{14}$$

where k is the size of the convolution kernel, s is the step size, p is the padding, and c_{out} is the number of output channels.

(4) Feed-forward layer: It increases non-linearity, improving the expressive ability of the model:

$$x_{ffn} = FeedForward(y_t; \omega_1, \sigma, \omega_2) \tag{15}$$

where w_1 and w_2 are parameters.

Finally, the long-term dependence representation X'_{rep} output by the *MoA* module can be obtained as:

$$X'_{rep} = MoA(X_{rep}) \tag{16}$$

Output of the encoder module: The output of the encoder module is the feature representation X'_{rep} generated by combining the *MoF* and *MoA* modules, which is fed into the subsequent prediction layer. Finally, the prediction result is obtained by the model through a linear projection:

$$\hat{X} = P(Flatten(X'_{rep})) \tag{17}$$

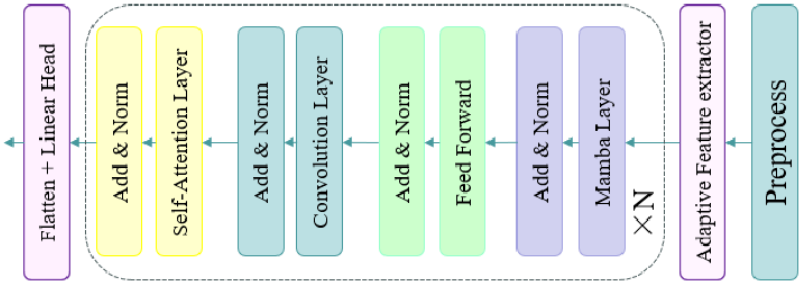

Figure 5. Details of the MoA.

3.4. Decoder Module

In the decoding stage, the future multi-modal motion distribution is explicitly modeled by the model through the introduction of intention anchors. Intention anchors are a set of learnable query vectors used to characterize potential different motion patterns; each anchor queries the encoded historical trajectory context during the decoding process, and is updated and refined layer by layer in the multi-layer decoder to obtain a more accurate depiction of the multi-solution future trajectory.

To improve the rationality of anchor initialization, the initial positions of intention anchors are set based on the clustering results of the trajectory distribution in the training data, ensuring they cover the main motion pattern clusters. Subsequently, the pattern representation of specific samples is gradually converged to by the anchors during the decoder's iterative update. The initial embedding of the anchor is as follows:

$$M_0 = MLP(A) + PE_A \tag{18}$$

where A represents the intention anchor parameters, and PE_A is the positional encoding related to the anchor position, used to inject structured position information.

Iterative Refinement Module (IRM): To refine the representation of trajectory intention anchors layer by layer and enhance the model's ability to model complex spatiotemporal dependence, an Iterative Refinement Module (IRM) is designed in the decoder. Taking learnable intention anchors as Query, and the historical trajectory context output by the encoder as Key and Value, this module gradually refines the semantic representation of anchors through a multi-layer Transformer structure, realizing dynamic alignment from the global modal prior to the sample conditional modal [31]. The iterative refinement

module stacks *N*-layer structures, and each layer includes three stages: Self-Attention, Cross-Attention, and Feed-Forward transformation. The input of the *j* layer is the output of the M^{j-1} layer, and its update process is as follows:

(1) Self-Attention layer (Anchor-to-Anchor Attention)

First, it models the interior of the anchors, enabling interactions between different modalities and capturing potential correlation relationships:

$$C_{sa}^{j} = MultiHead\left(Q = M^{j-1}, K = M^{j-1}, V = M^{j-1}\right) \tag{19}$$

where the multi-head attention mechanism is used to model the dependence structure of different subspaces in parallel, and the output of each head is spliced and linearly transformed to fuse into a global representation. The problem of modal independence between anchors is effectively alleviated, and the complementarity of multi-modal features is improved by this step.

(2) Cross-Attention layer (Mode-to-History Attention)

The anchor features after self-attention are used as queries to select the spatiotemporal information most relevant to the current modality from the historical context *H*, realizing cross-modal feature alignment:

$$C_{ca}^{j} = MultiHead\left(Q = C_{sa}^{j}, K = H, V = H\right) \tag{20}$$

Through this interaction, the model can adaptively extract differentiated evidence from historical trajectories under different intentions, thereby refining the current pedestrian's intention layer by layer.

(3) Feature Transformation layer

The anchor update features C_{ca}^{j} are further nonlinearly mapped through the feed-forward network to obtain the output of the *j* layer:

$$M^{j} = FFN\left(C_{ca}^{j}\right) \tag{21}$$

Moreover, the training process is stabilized through residual connection and layer normalization (Add&Norm). The output of each layer, M^{j}, serves as the input to the next layer, enabling layer-by-layer semantic refinement and intention alignment. To strengthen multi-layer supervision and training stability, a trajectory regression branch Y^{j}, is connected after each layer to generate multi-modal future trajectories. The final prediction result is obtained by weighted fusion of the output of each layer:

$$\hat{Y} = \sum_{j=1}^{N} a_j Y^{j} \tag{22}$$

where the weight coefficient is set as the mean value or learnable parameters to balance the contribution of different layers.

4. Experiments

To fully verify the innovation and effectiveness of the proposed MUGI-Net, this chapter designs a systematic experimental scheme on two benchmark datasets. First, the experimental datasets, evaluation metrics, and implementation details are unified to ensure fair comparison. The overall performance of MUGI-Net is then compared with state-of-the-art methods to verify its superiority. Finally, ablation studies are conducted to evaluate the core modules.

4.1. Datasets

The proposed method is evaluated on two benchmarking datasets: JAAD [36] and PIE [37]. The JAAD dataset contains 686 pedestrians collected by vehicle front cameras at 30 Hz; the PIE dataset contains 1842

pedestrians with action annotations, also collected at 30 Hz. Video frames, pixel-level pedestrian coordinates and bounding box positions, and action/motion information for pedestrians and ego-vehicles are provided by both datasets. For pedestrian behavior annotation, both datasets use a consistent coding scheme that divides pedestrian actions into "walking" and "standing". The annotation of ego-vehicle motion state is slightly different in the two datasets: the ego-vehicle actions in JAAD are given in the form of semantic labels such as "deceleration, acceleration, fast movement, slow movement", while PIE provides continuous velocity and yaw angle measured by on-board sensors (such as gyroscopes). To achieve consistent modeling across datasets, the ego-vehicle states in JAAD are recoded into four discrete states: 0—stop, 1—slow driving, 2—fast driving, and 3—deceleration. Existing works [21,38] are referred to for data division and sequence settings to ensure fair comparisons: a historical sequence with a length of 0.5 s is used for different modal inputs (*i.e.*, under a 30 Hz sampling rate, the length of the historical sequence is set to $T_h = 15$ frames), and the prediction horizons are set to 0.5 s, 1.0 s and 1.5 s respectively, corresponding to the prediction sequence lengths of 15, 30, and 45 frames.

4.2. Evaluation Metrics

To accurately measure the prediction error of pedestrian trajectories in FPV (characterized by bounding box scale changes), three quantitative metrics are adopted based on the displacement between predicted and ground-truth trajectories. Based on established works [21,36,37], our method uses: (1) Bounding Box Mean Square Error (MSE) of the top-left and bottom-right coordinates; (2) Center Mean Square Error (CMSE) of the bounding box; (3) Center Final Mean Square Error (CFMSE) of the center point coordinates of the bounding box, as follows:

$$\mathrm{MSE} = \frac{1}{T_f} \sum_{t=1}^{T_f} \| \hat{Y}_t - Y_t \|_2^2 \tag{23}$$

$$\mathrm{CMSE} = \frac{1}{T_f} \sum_{t=1}^{T_f} \| \hat{c}_t - c_t \|_2^2 \tag{24}$$

$$\mathrm{CFMSE} = \| \hat{c}T_f - cT_f \|_2^2 \tag{25}$$

where $\hat{Y}_t$ and Y_t represent the ground-truth and predicted bounding boxes at time step t, $\hat{c}_t$ and c_t represent the corresponding ground-truth and predicted center coordinates.

4.3. Implementation Details

The model is trained on a single RTX 3060 GPU, using the Adam optimizer with an initial learning rate of 0.001, and the learning rate is adjusted based on the validation set loss with a decay factor of 0.2. The dimension of the hidden layer is set to $d_h = 256$, and the dimension of the input embedding is set to $d_0 = 128$. The number of layers in the Transformer encoder is $N = 3$. The feature dimensions of pedestrian trajectory and pedestrian motion are $N_x = 4$ and $N_p = 1$, respectively. The feature dimension of ego-vehicle motion differs between the JAAD and PIE datasets: 1 and 2, respectively. In addition, the number of heads in the attention mechanism is set to $h = 8$, and the performance is evaluated based on the modality with the best results ($K = 20$). The training is carried out for 50 epochs, with a batch size of 128 and a dropout rate of 0.1.

4.4. Quantitative Analysis

To verify the overall performance of MUGI-Net (the integrated effect of all core modules), this section compares the model with 7 state-of-the-art baseline methods (SGNet (2022) [18], Adsampler (2024) [39],

AMTN (2024 b1) [39], OS-TF-S (2024) [40], SGNetPose (2025) [41], ABC+ (2023) [42], and AANet (2025) [22]) on JAAD and PIE datasets, with the results shown in Table 2 (bold indicates the best performance).

On the JAAD dataset (1.5 s medium-long term prediction, MSE_15), MUGI-Net achieves 153, a 5% reduction compared with the SOTA method AANet (161), and outperforms all other baselines by a significant margin; in CFMSE (343), the model also surpasses AANet (356), verifying that the integrated design of group interaction and MoU temporal encoding effectively reduces the long-term prediction error accumulation caused by insufficient interaction modeling and temporal dependence capture.

On the PIE dataset (1.0 s medium-term prediction, MSE_10), MUGI-Net attains the optimal value of 33 (on par with ENCORE-NR). In MSE_15 (73), CMSE (52), and CFMSE (140), it is significantly superior to AANet and most baselines, confirming that the model exhibits stable, excellent performance across different datasets. The core modules exhibit strong generalization across FPV scenarios.

Even in short-term prediction (0.5 s, MSE_05), MUGI-Net (34 on JAAD, 16 on PIE) is only slightly lower than the optimal method ENCORE-NR and better than all other baselines, indicating that the model does not sacrifice short-term prediction accuracy while optimizing medium and long-term performance, which is the direct effect of the MoU structure's balanced capture of short-term dynamics and long-term dependence.

Table 2. Experimental results on JAAD and PIE datasets (Bold indicates the best performance).

	JAAD					**PIE**				
	MSE_05	**MSE_10**	**MSE_15**	**CMSE**	**CFMSE**	**MSE_05**	**MSE_10**	**MSE_15**	**CMSE**	**CFMSE**
SGNet (2022)	37	86	197	146	443	**16**	39	88	66	206
Adsampler (2024)	42	88	175	127	**322**	**16**	38	77	54	**133**
AMTN (2024)	38	81	179	133	337	17	38	79	55	143
OS-TF-S (2024)	40	90	184	142	395	18	39	82	62	172
SGNetPose (2025)	62	146	347	260	872	**16**	40	103	80	272
ABC+ (2023)	40	89	189	145	409	**16**	38	87	65	191
AANet (2025)	35	**74**	161	**121**	356	**16**	35	78	53	160
Ours	**34**	78	**153**	130	343	**16**	**33**	**73**	**52**	140

4.5. Qualitative Analysis

The qualitative evaluation results of our model on the JAAD and PIE datasets are shown in Figures 6 and 7. Our method achieves excellent predictive accuracy, especially in scenarios where pedestrians may exhibit multiple behavior patterns. As shown in Figures 6 and 7, the model generates future trajectories across different urban scenarios and accurately captures diverse motion patterns. In the figures, blue represents the observed trajectory, red represents the ground truth trajectory, and green represents the predicted trajectory. For example, in the middle subgraph, the model successfully predicts different future motions, such as crossing the road or continuing forward. This is the joint effect of the group interaction module (capturing social interaction) and the intention anchor decoder (generating multi-modal, reasonable trajectories), verifying that the model can generate trajectories that conform to real traffic social norms rather than just pursuing numerical error reduction.

Figure 6. Qualitative analysis on JAAD dataset.

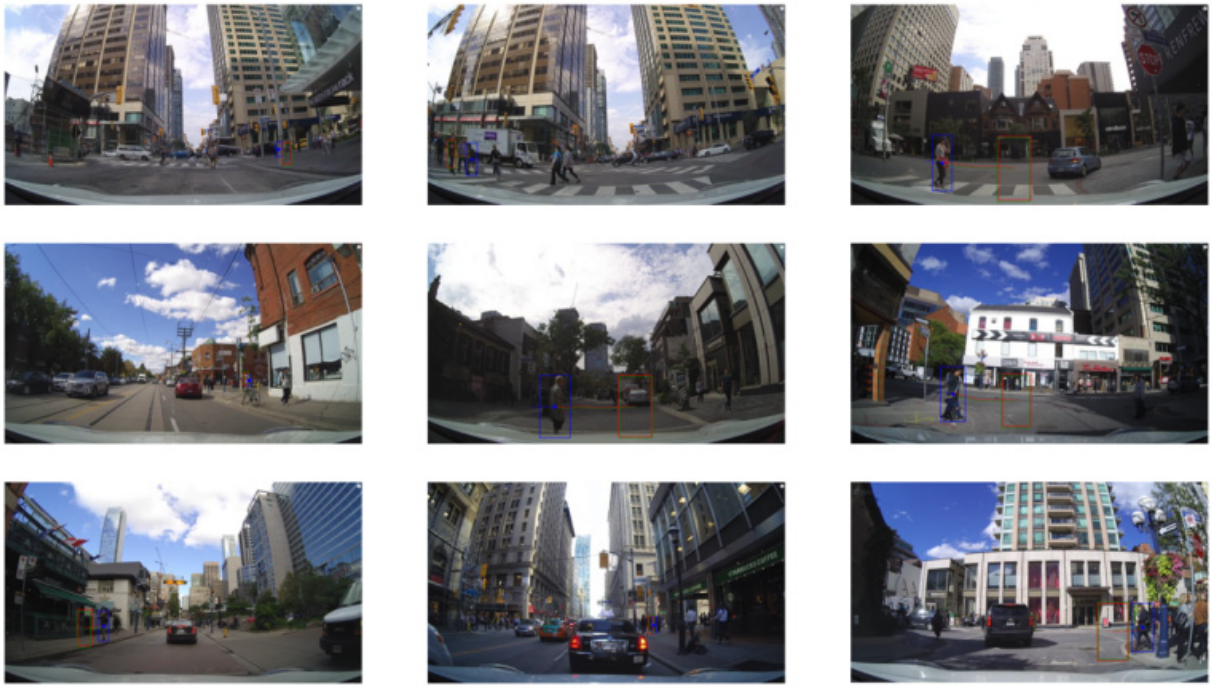

Figure 7. Qualitative analysis on PIE dataset.

4.6. Ablation Study

To verify the design's effectiveness, a comprehensive ablation experiment is conducted on the group interaction and temporal encoding modules (*i.e.*, the first-stage encoder). For the part without group interaction, only the historical bounding box is input. For the part without group interaction, the only input is the historical bounding box. Tables 3 and 4 show the results obtained from the JAAD and PIE datasets. In Table 3, the symbol √ indicates the use of Group Interaction, while the symbol × indicates the absence of Group Interaction.

Table 3. Impact of group interaction on pedestrian trajectory prediction (Bold indicates the best performance).

	JAAD					PIE				
	MSE			CMSE	CFMSE	MSE			CMSE	CFMSE
Group Interaction	0.5 s	1.0 s	1.5 s			0.5 s	1.0 s	1.5 s		
√	**34**	**78**	**153**	**130**	**343**	**16**	**33**	**73**	**52**	**140**
×	38	83	178	156	367	18	38	79	60	179

Table 4. Impact of different encoding methods on pedestrian trajectory prediction (Bold indicates the best performance).

	JAAD					PIE				
	MSE			CMSE	CFMSE	MSE			CMSE	CFMSE
Temporal Encoding Method	0.5 s	1.0 s	1.5 s			0.5 s	1.0 s	1.5 s		
MoU (Ours)	**34**	**78**	**153**	**130**	**343**	**16**	**33**	**73**	**52**	**140**
Transformer	36	79	165	149	382	18	36	79	69	179
Mamba	38	89	158	131	397	17	38	83	63	185

4.6.1. Impact of Group Interaction

The impact of group information on pedestrian trajectory prediction is illustrated in Table 3. After adding group information, trajectories are predicted by the model based on distance and direction information among pedestrians. Figures 8 and 9 show the prediction results with and without the group interaction module for the JAAD and PIE datasets, where blue represents the observed trajectory, red represents the ground truth trajectory, green represents the predicted trajectory with the group interaction module, and yellow represents the predicted trajectory without it. The superiority of the prediction results with the group interaction module over those without it is shown by the figures. Therefore, the relevant performance indicators on the JAAD/PIE datasets are significantly improved.

Figure 8. Experimental comparison with and without group interaction module on JAAD dataset.

Figure 9. Experimental comparison with and without group interaction module on PIE dataset.

4.6.2. Impact of Temporal Encoding

The impact of different temporal encoding methods on pedestrian trajectory prediction performance is illustrated in Table 4. Since only long- or short-term dependence is often focused on by Transformer or Mamba encoders, a decline in prediction performance is caused by using them separately. Figures 10 and 11 show the results of different encoding methods in the JAAD and PIE datasets, where blue represents the observed trajectory, red represents the ground truth trajectory, green represents the predicted trajectory using the MoU encoding method, and yellow represents the predicted trajectory using the Transformer encoding method. The fact that superior predictions are yielded by the MoU encoding method compared to the Transformer encoding method is shown by the figures. Therefore, the relevant performance indicators on the JAAD/PIE datasets are significantly improved. This directly verifies that the MoF (adaptive short-term feature extraction via sliding window and sub-extractors) and MoA (hierarchical fusion of Mamba/Convolution/Self-Attention for global-long term modeling) structure effectively balances the capture of short-term dynamic details and long-term temporal dependence, making up for the inherent defects of Transformer (weak local detail capture, high computation) and Mamba (easily lost information in long-term prediction).

Figure 10. Visualization results of different temporal encoding methods on JAAD dataset.

Figure 11. Visualization results of different temporal encoding methods on PIE dataset.

4.6.3. Impact of the Number of Predicted Trajectories

The impact of different numbers of predicted trajectories on pedestrian trajectory prediction performance is reported in Table 5. Since fewer predicted trajectories limit the model's ability to explore diverse plausible motion patterns and capture the inherent uncertainty of pedestrian behavior, prediction performance declines when the number of predicted trajectories is reduced from 20 to 10 or 5. The table presents results for different numbers of predicted trajectories in the JAAD and PIE datasets, where 20 (Ours) denotes our method with 20 predicted trajectories, and 10/5 denotes baseline settings with fewer trajectories. The table shows that superior performance (lower MSE/CMSE/CFMSE values across all time steps) is achieved with 20 predicted trajectories compared to 10 or 5. Therefore, the relevant performance indicators on the JAAD/PIE datasets are significantly improved when using 20 predicted trajectories. This directly verifies that a larger number of predicted trajectories effectively enhances the model's capacity to capture multi-modal pedestrian motion patterns and reduce prediction error, making up for the limitations of fewer trajectories in exploring the full range of possible future movements and handling uncertainty in pedestrian behavior.

Table 5. Impact of the number of predicted trajectories on pedestrian trajectory prediction (Bold indicates the best performance).

	JAAD					**PIE**				
	MSE			**CMSE**	**CFMSE**	**MSE**			**CMSE**	**CFMSE**
Number of Predicted Trajectories	**0.5 s**	**1.0 s**	**1.5 s**			**0.5 s**	**1.0 s**	**1.5 s**		
20 (Ours)	**34**	**78**	**153**	**130**	**343**	**16**	**33**	**73**	**52**	**140**
10	51	147	400	354	1051	18	46	114	91	311
5	83	357	1148	1110	3328	27	88	258	234	873

5. Conclusions

This work proposes MUGI-Net, a FPV pedestrian trajectory prediction model that fuses group interaction and hybrid temporal encoding, addressing deficiencies in existing methods for group interaction modeling and temporal dependence capture. On the JAAD dataset, the model achieves an MSE_15 of 153 for 1.5 s prediction, a 5% reduction compared with the SOTA AANet (161), and outperforms most baselines in CFMSE. On the PIE dataset, it attains an optimal MSE_10 of 33 for 1.0 s prediction and excels in MSE_15, CMSE, and CFMSE compared with mainstream methods such as AANet and SGNet. Ablation

experiments further confirm that the group interaction module reduces JAAD's MSE_15 by 14.04% and PIE's by 7.59%, while the MoU temporal structure outperforms single Transformer/Mamba encodings by 7.27% and 3.16% in JAAD's MSE_15, respectively. These results directly validate the innovation and effectiveness of the proposed group pooling mechanism, MoU hybrid temporal encoding, and intention anchor-based iterative refinement: the group interaction module realizes fine-grained modeling of pedestrian social relationships for FPV scenarios, the MoU structure balances short-term dynamic and long-term dependence capture, and the decoder boosts multi-modal prediction stability. MUGI-Net significantly improves the medium- and long-term prediction accuracy of FPV pedestrian trajectories, provides an effective technical solution for autonomous driving in complex mixed traffic scenarios, and enriches the theoretical and methodological framework of FPV-based pedestrian trajectory prediction, with significant research value for advancing the practical application of autonomous driving and intelligent transportation systems.

6. Discussion

MUGI-Net's superior performance on the JAAD and PIE datasets verifies that group interaction modeling and a MoU-based hybrid temporal encoding are critical for FPV pedestrian trajectory prediction. The group interaction module, via dynamic grouping and sparse intra- and inter-group graphs, effectively captures pedestrian social interactions and mitigates FPV perspective drift, reducing medium- to long-term prediction errors by over 7% across key metrics. The MoU structure balances short-term dynamic capture and long-term dependence, outperforming single Transformer/Mamba encodings by avoiding detail loss and computational inefficiency. Intention anchor-based IRM enhances the rationality of multi-modal prediction, adapting to FPV's motion randomness. However, the model is limited in severe occlusion scenarios and in complex environmental integration, and its inference speed needs to be optimized for real-time autonomous driving. Future research will focus on multi-sensor fusion, environmental factor modeling, and model lightweighting to further improve its practical application value in autonomous driving systems.

Statement of the Use of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the author(s) used Deepseek for grammer polishing. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Acknowledgments

This work is supported by General Project of National Natural Science Foundation of China NO.62576052, General Project of Jiangsu Provincial Department of Science and Technology NO.BK20250969, Changzhou City Applied Basic Project No.CJ20240039, and Changzhou Leading Innovative Talent Introduction and Cultivation Project NO.CQ20250044.

Author Contributions

Conceptualization, R.N. and B.Y.; Methodology, R.N.; Software, M.D.; Validation, S.Y. and M.D.; Formal Analysis, S.Y.; Investigation, M.D.; Resources, B.Y.; Data Curation, M.D.; Writing—Original Draft Preparation, R.N.; Writing—Review & Editing, B.Y.; Visualization, M.D.; Supervision, B.Y.; Project Administration, B.Y.; Funding Acquisition, R.N. and B.Y.

Ethics Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The research data is based on publicly available datasets on the internet.

Funding

This research was funded by General Project of National Natural Science Foundation of China under grant number 62576052; Project of Jiangsu Provincial Department of Science and Technology under grant number BK20250969; Changzhou City Applied Basic Project under grant number KYZ24020273; Changzhou Leading Innovative Talent Introduction and Cultivation Project under grant number CQ20250044.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Golchoubian M, Ghafurian M, Dautenhahn K, Azad NL. Pedestrian Trajectory Prediction in Pedestrian-Vehicle Mixed Environments: A Systematic Review. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 11544–11567. DOI:10.1109/TITS.2023.3291196
2. Chen H, Liu Y, Hu C, Zhang X. Vulnerable Road User Trajectory Prediction for Autonomous Driving Using a Data-Driven Integrated Approach. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 7306–7317. DOI:10.1109/TITS.2023.3254809
3. Bai J, Fang X, Fang J, Xue J, Yuan C. Deep Virtual-to-Real Distillation for Pedestrian Crossing Prediction. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; pp. 1586–1592.
4. Chen W, Sang H, Wang J, Zhao Z. DSTIGCN: Deformable Spatial-Temporal Interaction Graph Convolution Network for Pedestrian Trajectory Prediction. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 6923–6935. DOI:10.1109/TITS.2024.3525080
5. Yang D, Zhang H, Yurtsever E, Redmill KA, Ozguner U. Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention. *IEEE Trans. Intell. Veh.* **2022**, *7*, 221–230. DOI:10.1109/TIV.2022.3162719
6. Chen W, Sang H, Wang J, Zhao Z. DSTCNN: Deformable Spatial-Temporal Convolutional Neural Network for Pedestrian Trajectory Prediction. *Inf. Sci.* **2024**, *666*, 120455. DOI:10.1016/j.ins.2024.120455
7. Chen X, Zhang H, Deng F, Liang J, Yang J. Stochastic Non-Autoregressive Transformer-Based Multi-Modal Pedestrian Trajectory Prediction for Intelligent Vehicles. *IEEE Trans. Intell. Transp. Syst.* **2023**, *25*, 3561–3574. DOI:10.1109/TITS.2023.3342040
8. Fang F, Wang X, Li Z, Qian K, Zhou B. A Unified Framework for Pedestrian Trajectory Prediction and Social-Friendly Navigation. *IEEE Trans. Ind. Electron.* **2023**, *71*, 11072–11082. DOI:10.1109/TIE.2023.3342301
9. Sharma N, Dhiman C, Indu S. Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration. *IEEE Trans. Intell. Veh.* **2024**, *9*, 6960–6970. DOI:10.1109/TIV.2024.3391898
10. Feng A, Han C, Gong J, Yi Y, Qiu R, Cheng Y. Multi-Scale Learnable Gabor Transform for Pedestrian Trajectory Prediction from Different Perspectives. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 13253–13263. DOI:10.1109/TITS.2024.3421373
11. Huang R, Ding J, Pagnucco M, Song Y. Fully Decoupling Trajectory and Scene Encoding for Lightweight Heatmap-Oriented Trajectory Prediction. *IEEE Robot. Autom. Lett.* **2024**, *9*, 9143–9150. DOI:10.1109/LRA.2024.3426376
12. Yao Y, Atkins E, Roberson MJ, Vasudevan R, Du X. Coupling Intent and Action for Pedestrian Crossing Behavior Prediction. *arXiv* **2021**, arXiv:2105.04133. DOI:10.48550/arXiv.2105.04133
13. Fu M, Zhang T, Song W, Yang Y, Wang M. Trajectory Prediction-Based Local Spatio-Temporal Navigation Map for Autonomous Driving in Dynamic Highway Environments. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 6418–6429. DOI:10.1109/TITS.2021.3057110

14. Phan-Minh T, Grigore EC, Boulton FA, Beijbom O, Wolff EM. CoverNet: Multimodal Behavior Prediction Using Trajectory Sets. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 14074–14083.
15. Dendorfer P, Elflein S, Leal-Taixe L. MG-GAN: A Multi-Generator Model Preventing Out-of-Distribution Samples in Pedestrian Trajectory Prediction. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 13158–13167.
16. Su Y, Du J, Li Y, Li X, Liang R, Hua Z, et al. Trajectory Forecasting Based on Prior-Aware Directed Graph Convolutional Neural Network. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 16773–16785. DOI:10.1109/TITS.2022.3142248
17. Yao Y, Atkins E, Johnson-Roberson M, Vasudevan R, Du X. BiTraP: Bi-Directional Pedestrian Trajectory Prediction with Multi-Modal Goal Estimation. *IEEE Robot. Autom. Lett.* **2021**, *6*, 1463–1470. DOI:10.1109/LRA.2021.3056339
18. Wang C, Wang Y, Xu M, Crandall DJ. Stepwise Goal-Driven Networks for Trajectory Prediction. *IEEE Robot. Autom. Lett.* **2022**, *7*, 2716–2723. DOI:10.1109/LRA.2022.3145090
19. Yin Z, Liu R, Xiong Z, Yuan Z. Multimodal Transformer Networks for Pedestrian Trajectory Prediction. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), Montreal, QC, Canada, 19–26 August 2021; pp. 1259–1265.
20. Pang Y, Zhao X, Hu J, Yan H, Liu Y. Bayesian Spatio-Temporal Graph Transformer Network (B-STAR) for Multi-Aircraft Trajectory Prediction. *Knowl.-Based Syst.* **2022**, *249*, 108998. DOI:10.1016/j.knosys.2022.108998
21. Xu C, Mao W, Zhang W, Chen S. Remember Intentions: Retrospective-Memory-Based Trajectory Prediction. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 6488–6497.
22. Niu R, Hu C, Yang B, Chen H, Dai Z. Egocentric Pedestrian Trajectory Prediction with Agent-Wise Motion Fusion for Internet of Vehicles. *IEEE Internet Things J.* **2025**, *12*, 37659–37669. DOI:10.1109/JIOT.2025.3583722
23. Yang B, Fan F, Ni R, Wang H, Jafaripournimchahi A, Hu H. A Multi-Task Learning Network with a Collision-Aware Graph Transformer for Traffic-Agents Trajectory Prediction. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 6677–6690. DOI:10.1109/TITS.2023.3345296
24. Yang B, Yan K, Hu C, Hu H, Yu Z, Ni R. Dynamic Subclass-Balancing Contrastive Learning for Long-Tail Pedestrian Trajectory Prediction with Progressive Refinement. *IEEE Trans. Autom. Sci. Eng.* **2024**, *22*, 8645–8658. DOI:10.1109/TASE.2024.3487255
25. Ni R, Fang L, Hu C, Cai Y, Hu H, Yang B. Privacy-Aware Pedestrian Trajectory Prediction with Dual-Channel Destination Guidance and Multi-Factor Aggregation Federated Learning. *Transp. B Transp. Dyn.* **2025**, *13*, 2601598. DOI:10.1080/21680566.2025.2601598
26. Yang B, He C, Wang P, Chan CY, Liu X, Chen Y. TPPO: A Novel Trajectory Predictor with Pseudo Oracle. *IEEE Trans. Syst. Man Cybern. Syst.* **2024**, *54*, 2846–2859. DOI:10.1109/TSMC.2024.3351859
27. Yuan Y, Weng X, Ou Y, Kitani K. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9793–9803.
28. Sun J, Jiang Q, Lu C. Recursive Social Behavior Graph for Trajectory Prediction. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 657–666.
29. Liu Y, Guo H, Meng Q, Chen H. Context-Aware Heterogeneous Fusion for Pedestrian Trajectory Prediction in Pedestrian-Vehicle Interactions. *IEEE Sens. J.* **2025**, *25*, 31040–31052. DOI:10.1109/JSEN.2025.3583223
30. Bae I, Park JH, Jeon HG. Learning Pedestrian Group Representations for Multi-Modal Trajectory Prediction. In *European Conference on Computer Vision*; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 270–289.
31. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. DOI:10.1109/TNNLS.2020.2978386
32. Ni R, Lu S, Hu C, Yang B. Adaptive Progressive Transformer-Based Trajectory Prediction under Fine-Grained Trajectory-Scene Interaction Constraint. *IEEE Trans. Autom. Sci. Eng.* **2025**, *22*, 24498–24509. DOI:10.1109/TASE.2025.3633415
33. Yang B, Lu Y, Wan R, Hu H, Yang C, Ni R. Meta-IRLSOT++: A Meta-Inverse Reinforcement Learning Method for Fast Adaptation of Trajectory Prediction Networks. *Expert Syst. Appl.* **2024**, *240*, 122499. DOI:10.1016/j.eswa.2024.122499
34. Yang B, Zhu J, Yu Z, Fan F, Liu X, Ni R. Fast Adaptation Trajectory Prediction Method Based on Online Multisource Transfer Learning. *IEEE Trans. Autom. Sci. Eng.* **2025**, *22*, 1289–1304. DOI:10.1109/TASE.2024.3362980
35. Cangea C, Veličković P, Jovanović N, Kipf T, Liò P. Towards Sparse Hierarchical Graph Classifiers. *arXiv* **2018**, arXiv:1811.01287. DOI:10.48550/arXiv.1811.01287
36. Gao H, Ji S. Graph U-Nets. In *International Conference on Machine Learning*; PMLR: Long Beach, CA, USA, 2019; pp. 2083–2092.

37. Rasouli A, Kotseruba I, Tsotsos JK. Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 206–213.
38. Rasouli A, Kotseruba I, Kunic T, Tsotsos J. PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6262–6271.
39. Liu Q, Sang H, Wang J, Chen W, Liu Y. Non-Probability Sampling Network Based on Anomaly Pedestrian Trajectory Discrimination for Pedestrian Trajectory Prediction. *Image Vis. Comput.* **2024**, *143*, 104954. DOI:10.1016/j.imavis.2024.104954
40. Lin Y, Hu C, Zhao B, Jiang H, Shan Y, Ding T, et al. Anchor-Based Multi-Modal Transformer Network for Pedestrian Trajectory and Intention Prediction. In Proceedings of the 2023 7th CAA International Conference on Vehicular Control and Intelligence (CVCI), Changsha, China, 27–29 October 2023; pp. 1–6.
41. Wang J, Sang H, Chen W, Zhao Z. VOSTN: Variational One-Shot Transformer Network for Pedestrian Trajectory Prediction. *Phys. Scr.* **2024**, *99*, 026002. DOI:10.1088/1402-4896/ad19b7
42. Ghiya A, AlShami AK, Kalita J. SGNetPose+: Stepwise Goal-Driven Networks with Pose Information for Trajectory Prediction in Autonomous Driving. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Tucson, AZ, USA, 28 February–4 March 2025; pp. 629–637.