DPIU: Dynamic Pedestrian Intention Understanding Through Cognitive Decision-Making

Jiaheng Xiao, Zhihui Li, Mingxin Wang, Yu Xie, Qin Ma, Xin Wang, and Yu Sun

***Abstract*—Accurate prediction of pedestrian motion is crucial for autonomous driving, particularly in path planning and collision avoidance applications. Most current methods concentrate on spatiotemporal feature parameters (e.g., velocity continuity, social force parameters, and poses) extracted from historical trajectories to model pedestrian movement. However, these methodologies fail to adequately capture pedestrian intent and do not dynamically account for behavioral heterogeneity, leading to significant discrepancies with real-world observations. To address this issue, a dynamic pedestrian intention understanding (DPIU) framework is proposed, which links future intentions to historical experiences. Grounded in cognitive decision-making mechanisms derived from human physiology, the DPIU framework is designed to predict pedestrian motion by inferring inherent movement intentions. To establish a comprehensive historical perspective, a multiscale detail feature module is employed, incorporating a time-scale-based trajectory segmentation strategy to enhance the representation of pedestrian states. Subsequently, a goal intent prediction module is introduced, employing a probabilistic model to estimate pedestrians' inclination toward the spatial scope of their intended goals. This module assesses the similarity between the current scene and historical experiences, thereby optimizing the utilization of time-fragmented information. Finally, a dynamic optimization module is developed, which superimposes intent point probabilities and applies a Bayesian-based density estimation method to ensure that the predicted outcomes closely align with real-world behaviors. Experimental evaluations on the Stanford drone drones (SDD), ETH-UCY, and ApolloScape datasets demonstrate that the proposed DPIU framework outperforms existing methods in predicting future trajectories and optimizing multimodal forecasting outcomes. The method substantially improves predictive performance in dynamic scenarios, providing a valuable tool for autonomous driving applications.**

***Index Terms*—Behavioral sciences, feature extraction, human physiology, memory architecture, pedestrian intention, predictive models, trajectory.**

I. Introduction

ACCURATE prediction of pedestrian motion is critical for autonomous driving, especially in applications such as smooth path planning [1] and collision avoidance [2], which present significant challenges for autonomous driving systems. This necessity becomes particularly critical in the context of autonomous vehicles and human-robot interactions [3]. The intentions of pedestrians, which govern their mobility, play a pivotal role in goal selection and the prediction of future trajectories. Inaccurate understanding of these intentions can result in misjudgments of safety levels [4], erroneous driving decisions, and compromised driving smoothness [3]. Such miscalculations can lead to traffic accidents and hinder adaptability in mixed pedestrian-vehicle scenarios [5]. Therefore, accurate prediction of pedestrian intentions is essential for ensuring vehicular safety and enhancing driving smoothness in autonomous driving systems.

Contemporary research primarily utilizes recurrent neural networks (RNNs) [6], [7], transformer [8], [9] and graph convolution neural networks (GCNs) [10], [11], [12] to infer pedestrian motion intentions based on features such as motion vectors [13], contours, and poses [14]. Although these methods focus on modeling pedestrian motion through feature parameters, they fail to capture the complex nature of pedestrian cognitive decision-making, resulting in significant deviations from real-world behavior.

Rooted in human physiology, shown in Fig. 1, behavior inherently follows goal-oriented principles [15], with individuals carrying goal intentions during motion that are influenced by historical experiences and significantly shape their short-term actions. The replication of historical choices is a common phenomenon, as human learning typically relies on the retrieval of historical information [16]. Moreover, intended motions depend on both the expectation of their successful completion and the degree of inclination toward the intended goal [17]. When the degree increases, the influence of intention on motion becomes more substantial. As pedestrians gradually approach their short-term destinations, their trajectories become more explicit. The model should be capable of accurately capturing these characteristics, allowing vehicles to maneuver around pedestrians without disrupting the overall traffic flow [18]. Also, external factors may cause pedestrian intentions to adapt to the current scene, resulting in deviations from previously similar goal intentions.

To capture nuanced motion patterns and enhance the realism of pedestrian intention prediction, we propose dynamic pedestrian intention understanding (DPIU), inspired by the cognitive decision-making mechanisms in human physiology, where behavior is shaped by historical experiences and driven by goal-directed intentions. DPIU predicts pedestrian intentions by establishing a connection between future goals and historical experiences, resulting in a more accurate rep-

Received 30 August 2024; revised 6 March 2025, 18 August 2025, and 12 October 2025; accepted 12 February 2026. This work was supported by the Science and Technology Development Plan Project of Jilin Province, China under Grant 20250203187SF. *(Corresponding authors: Yu Sun; Zhihui Li.)*

Jiaheng Xiao, Zhihui Li, Mingxin Wang, Yu Xie, Xin Wang, and Yu Sun are with Transportation College, Jilin University, Changchun 130015, China (e-mail: fighting pangyu@163.com; lizhih@jlu.edu.cn).

Qin Ma is with the School of Navigation, Jimei University, Xiamen 361021, China.

Digital Object Identifier 10.1109/TNNLS.2026.3665567

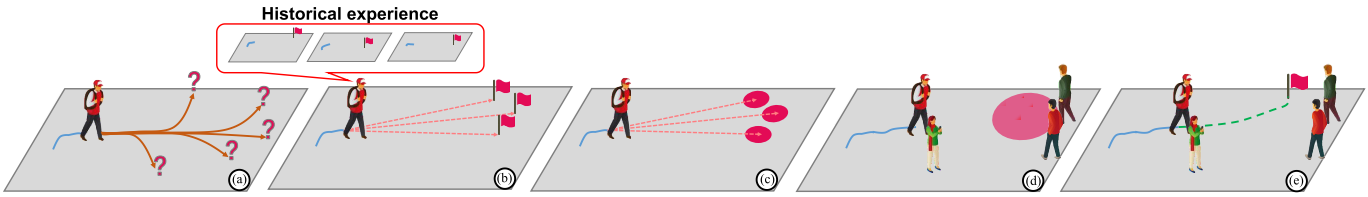

Fig. 1. Intention understanding process in DPIU. (a) Task Description: Given the past trajectory of the target pedestrian (depicted in blue), the objective is to predict the future trajectory within a few seconds. (b) Goal intentions are influenced by historical experience. (c) Incorporating the dynamic current scene causes regional-scale fluctuations in goal intent. (d) These fluctuations are examined to ensure the validity of the intention, while variations in regional scope and interaction environment affect goal intent selection during subsequent motion (Color shades indicate the extent of inclination.) (e) Future predicted trajectory is determined by the guidance of the goal intent.

resentation of pedestrian movements. The proposed DPIU framework consists of three key modules. First, to obtain a more comprehensive historical perspective, a multiscale detail feature module is employed. This module incorporates a time-scale-based trajectory segmentation strategy, which extracts time-fragmented information from past trajectories to enrich the representation of the pedestrian's state. Second, we introduce a goal intent prediction module that employs a probabilistic model to estimate the extent of a pedestrian's inclination toward the spatial domain of the intended goal. The model relies on assessing the similarity between the current scene and historical experiences. Since predefined functions (e.g., the cosine function) may not fully capture the similarity between feature vectors, a trainable addresser is introduced, incorporating a shallow attention network designed to learn similarity metrics. To maximize the use of time-fragmented information, weighting and filtering mechanisms are incorporated into the prediction process. Third, aligning with the realistic observation that pedestrians typically move toward a specific goal, a dynamic optimization module is incorporated, introducing Bayesian principles into the density estimation algorithm and strategically weighting the pedestrians' past trajectory tendencies with the multimodal prediction results. The integration of these elements enables the module to optimize multimodal predictions, enhancing their alignment with actual pedestrian behavior. Experimental evaluations on three datasets—Stanford drone drones (SDD), ETH-UCY, and ApolloScape—demonstrate the superiority of DPIU in predicting future trajectories accurately and optimizing multimodal forecasting results with minimal loss of predictive accuracy compared to previous state-of-the-art methods.

This article presents significant contributions as follows.

1) We propose DPIU, a novel pedestrian intention understanding network that captures the essence of cognitive decision-making and effectively leverages historical experience to enhance predictive performance in dynamic scenarios.
2) Three pivotal modules are proposed within DPIU: 1) multiscale detail feature module, which enhances the representation of past pedestrian motions through a time-scale-based trajectory segmentation strategy; 2) goal intent prediction module introduces a probabilistic model to explicitly quantify pedestrians' tendencies toward future intentions in a multimodal context; and 3) Bayesian principles are integrated into the dynamic optimization module to enhance the optimization of multimodal prediction results.
3) We empirically evaluate our approach on three extensive datasets, showing that it outperforms existing methods. Moreover, DPIU achieves more reliable optimization of multimodal prediction results with minimal loss of accuracy, demonstrating its suitability for autonomous driving applications.

The remainder of this article is organized as follows. Related work is presented in Section II. The framework and details of DPIU are described in Section III. Experiments and discussion are provided in Section IV. Finally, Section V concludes the article.

II. Related Work

With the rapid advancement of deep learning and growing interest in autonomous driving technologies, the field of pedestrian trajectory prediction has witnessed significant progress. Given the extensive research in this area, the focus of the present study is directed toward specific aspects, which are elaborated in detail in this section.

A. Parameter-Based Pedestrian Trajectory Prediction

Early research on trajectory prediction primarily focused on manual deterministic methods, employing hand-crafted rules and generic functions, such as Markov processes [4], to analyze pedestrian motion. Although these approaches aimed to improve the understanding of pedestrian dynamics and highlight the importance of social interactions, they struggled to capture critical features of pedestrian motion and showed limitations in handling complex scenarios.

In recent years, with the emergence of data-driven models, researchers have introduced multimodal trajectory prediction frameworks that employ parameter-based statistical models to represent behavioral patterns [19]. For instance, researchers such as [13] and [20] have utilized encoder–decoder structures to model pedestrian trajectory characteristics, while [21] has addressed the long-tail distribution in pedestrian trajectory prediction. Wong et al. [22] proposed SocialCircle, which quantifies interactions through the ternary components of speed, distance, and direction to enhance the prediction accuracy of multiple models. Moreover, some studies have employed conditional variational autoencoder (CVAE) [23], [24], and [25] to make predictions by estimating the parameters of an intermediate distribution and sampling future

features from it. Yang et al. [14] and Li et al. [26] have utilized a generative adversarial network (GAN) to learn motion patterns and generate future predictions. Some researchers have also explored future prediction distributions by utilizing a Gaussian mixture model [27]. However, these methods did not account for the movement direction of neighboring pedestrians or higher order interactions, and their ability to model complex scenarios was limited. Interaction enhancement methods, such as GTPPO [28] and TP-EGT [29], have improved trajectory prediction. GTPPO reduces the gap between historical and future knowledge by combining obstacle avoidance experiences with pseudo-prophecy machines through graph attention, whereas TP-EGT employs a collision-aware graph transformer to improve safety by predicting interaction probabilities. Meta-IRLSOT [30] addresses these issues by employing inverse reinforcement learning to align trajectories with scene semantics and leveraging meta-learning to enable rapid adaptation with limited data. However, both methods struggle to capture sudden changes in pedestrian behavior, such as sharp turns, and rely on fixed scene data, thereby limiting their adaptability to new environments.

B. Instance-Based Pedestrian Trajectory Prediction

While parameter-based methods can capture general motion patterns among pedestrians, they often overlook the diversity and variability in pedestrian motion, thereby masking individual differences. In addition, these methods struggle to provide a coherent cognitive framework for guiding future predictions. Addressing these limitations, recent research, such as [20], has integrated spatial constraints and conditional probabilities into future predictions, effectively managing the stochasticity of future trajectories and incorporating pedestrian intentions to enhance prediction accuracy. To gain a more comprehensive understanding of the subject matter and establish a direct link between the current scene, historical data, and individual pedestrian differences, some researchers have introduced example-based methodologies [31], [32], [33]. These methods emulate the retrospective memory processes in pedestrian decision-making, using historical instances to guide predictions of the current scenario. For instance, Marchetti et al. [32] effectively employed a memory mechanism to predict trajectories for individual agents, while [31] maintained memory bank pairs containing instances from the training data for relevant motion pattern matching during the prediction process.

Although these approaches touch on the essence of pedestrian cognitive decision-making, they focus mainly on motion features and offer limited insight into their underlying nature. Consequently, they are less suited for accurately capturing pedestrian intentions, particularly in cases involving complex deviations from standard motion patterns.

C. Pedestrian Observation and Decision-Making Mechanism

Within the field of physiology, prior research, exemplified by [16], has investigated the memory characteristics that shape human behavior. These findings suggest that decisions in a given scene are often informed by historical experiences that mirror the current context. Goldman-Rakic [34] explores the concept of inertia in human motion, highlighting that future movements are influenced by past motion, thus underscoring the role of historical motion in shaping future behavior. Nonetheless, it is essential to recognize that human motion inherently involves stochastic elements due to the inherent uncertainty [35], introduced by stochastic variables [36]. In addition, Helbing [37] acknowledges the stochastic nature of human motion, attributing it to various factors, such as the surrounding environment, pedestrians' walking habits, and interactive variables, including other pedestrians. It is important to recognize that human future motion is not simply a linear extrapolation of past actions but is profoundly influenced by the unconscious stochasticity inherent in human decision-making. Furthermore, Tomasello et al. [15], and Wolinski et al. [38] investigate how goal tendencies evolve over the course of human motion.

Expanding on the prior analysis, our model improves the ability to understand pedestrian intentions, facilitating accurate predictions of a wide range of motion behaviors. It simulates shifts in pedestrian intent by explicitly quantifying their predisposition toward future actions. In addition, we leverage Bayesian concepts to enhance the accuracy of future predictions. In conclusion, our proposed DPIU surpasses the model described in [31] across the SDD, ETH-UCY, and ApolloScape datasets.

III. Proposed Method

A. Problem Formulation

Trajectory prediction aims to predict the future trajectories of the target agent based on past trajectories and their neighboring agents. Let $x^t \in R^2$ be the spatial coordinate of the target agent at timestamp t, and $\mathbf{X} = [\mathbf{x}^{-T_p+1}, \mathbf{x}^{-T_p+2}, \ldots, \mathbf{x}^0] \in R^{T_p \times 2}$ be the past trajectory of the target agent at the observation timestamp; then $\mathcal{N}$ be the set of neighboring agents of target agent, and $\mathcal{X}_\mathcal{N} = [\mathbf{X}_{\mathcal{N}_1}, \mathbf{X}_{\mathcal{N}_2}, \ldots, \mathbf{X}_{\mathcal{N}_N}] \in R^{N \times T_p \times 2}$ denote the set of neighboring agents' past trajectories, where $X_{N_l} \in R^{T_p \times 2}$ is past trajectory of the lth neighboring agent. The future trajectory of target agent is denoted as $\hat{\mathbf{Y}} = [\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^{T_f}] \in R^{T_f \times 2}$, where $\mathbf{y}^t \in R^2$ is the spatial coordinate of the future timestamp t. The overall goal $G(\cdot)$ is to train the prediction model so that the future trajectories $\hat{\mathbf{Y}} = G(\mathbf{X}, \mathcal{X}_\mathcal{N})$ predicted by the model are as close as possible to the ground truth $\mathbf{Y}$, in addition, $\hat{\mathbf{y}}_{inten} = \hat{\mathbf{y}}^{T_f}$ is assumed that the predicted endpoint is the goal intent. The spatial coordinates themselves encode the velocity through the time difference (Δx/Δt), a method that avoids the noise amplification that is often associated with direct velocity estimation, resulting in a more robust and stable reflection of the pedestrian's movement pattern.

B. Structure Overview

To address our objective, we propose DPIU, depicted in the structural diagram presented in Fig. 2. The model comprises three key components: 1) a multiscale detail feature module; 2) a goal intent prediction module; and 3) a dynamic optimization module.

Recent research [34] in neuroscience suggests that pedestrian locomotion is not solely reactive, but is fundamentally grounded in a memory-based, hierarchical perceptual system

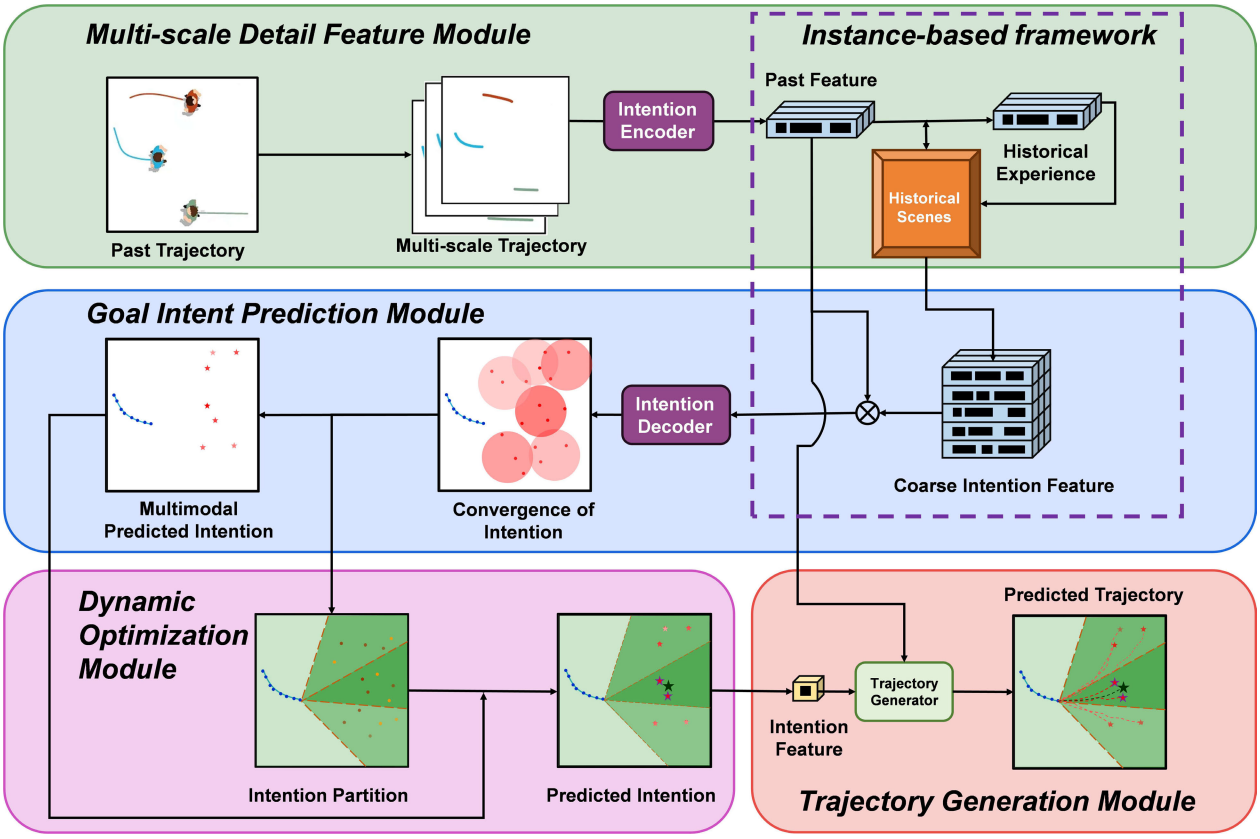

Fig. 2. Structure of the proposed model is illustrated as above. In the goal intent prediction module, the red dots represent coarse intentions, while the red stars denote predicted intentions. The color intensity reflects the probability estimate assigned, based on the similarity to the current trajectory stored in the memory bank, with darker shades indicating higher probability. In the dynamic optimization module, the red stars represent multimodal predicted intentions, the red stars with purple edges indicate optimized predicted intentions, and the dark green stars represent the ground truth.

[16], [39], [40]. Such a mechanism aligns fundamentally with the principle of predictive coding [41], which posits that the brain leverages prior knowledge to construct predictive models that are temporally refined via the continuous correction of prediction errors [42]. This iterative error correction allows for the dynamic integration of fragmented peripheral visual inputs, thereby constructing coherent environmental representations, ultimately accomplishing dynamic complementation of environmental representations under fragmented perception [43].

Inspired by the observed characteristics of human locomotion, the multiscale detail feature module utilizes the full-length trajectory history as core data, dividing it according to different time scales. This approach provides richer references for motion patterns and aids in identifying influential underlying factors. Moreover, the training analysis of past trajectories at different time scales is conducted to establish connections with historical experiences, enhancing generalization capabilities in the current scenario.

The goal intent prediction module plays a pivotal role in simulating human motion by analyzing the agent's motion tendencies toward goal intent and shifts in scene interaction. This module leverages historical experiences to construct goal intents, which are defined within a specific scope, thereby guiding the current scene. Given the diverse tendencies exhibited by agents, this variability is harnessed to emulate human decision-making processes. In addition, techniques for assigning information weights and eliminating redundant data are incorporated to enhance prediction accuracy.

The dynamic optimization module is designed to further refine prediction results, thereby achieving higher accuracy. The core of this module is a probabilistic model comprising two components: one uses the predicted outcomes of time-fragmented information as the prior probability, and the other employs the tendencies of predicted outcomes from full-length information as the posterior probability.

C. Multiscale Detail Feature Module

To enhance the precision of subtle motion capture and the expansion of relevant information, past trajectories are represented across multiple scales. This approach involves constructing a comprehensive representation of the current scene, extracting nuanced motions, and establishing connections between historical experiences and crucial data within the current scene. The ultimate goal of intent creates a more comprehensive representation of agents' past motion states.

Human observers frequently rely on fragmented information to avoid collisions and react to their surroundings, even when their observations are not continuous. Fragmented information aids in capturing subtle motion details, enhancing the effectiveness of avoiding sudden changes. In addition, human motion exhibits inertia, with the proximity of an agent's motion to the

predicted starting point significantly impacting future predictions. Therefore, time-fragmented information is instrumental in capturing fine-grained motion nuances, rendering it more suitable for trajectory analysis compared to full-length data. It is particularly adept at capturing abrupt motion changes, given that such changes constitute a larger proportion of time-fragmented information, which facilitates a more effective analysis of these details and their close relationship with subsequent predictions.

While time-fragmented information holds significance, full-length information provides a broader contextual understanding and a more comprehensive basis for guiding future predictions. Relying solely on time-fragmented information can lead to one-sided predictions. Therefore, we introduce time-fragmented information as a correction term, aimed at refining predictions derived from full-length information by discarding a portion of the full-length predictions and integrating the predictions from time-fragmented information.

Based on above thoughts, we define $\mathbf{X}^{full}$ as the full-length information of past trajectory, the last 3/4-length past trajectory is defined as time-fragmented information $\mathbf{X}^{quar}$ and half-length information is $\mathbf{X}^{half}$

$$\begin{aligned} \mathbf{X}^r &= \left[\mathbf{x}^{-T_{p_r}+1}, \mathbf{x}^{-T_{p_r}+2}, \ldots, \mathbf{x}^0\right] \\ r &\in (full, quar, half) \end{aligned} \tag{1}$$

where T_{p_r} is past timestamp. $\mathbf{X}_N$ is divided into $\mathbf{X}_N^r$, each of $\mathcal{X}_{\mathcal{N}}^r \in \{\mathbf{X}_{\mathcal{N}_1}^r, \mathbf{X}_{\mathcal{N}_2}^r, \ldots, \mathbf{X}_{\mathcal{N}_N}^r\}$ represents a set of past trajectories at different time scales.

The structure of the module is an instance-based framework, which focuses on selecting representative features, then saving them in historical scene units as historical experiences f_{his}^r through screening and establishing similarity metrics. Then, we extract the information in relevant historical experience through the connection between the current scene and historical experience, utilizing it to guide prediction. Past feature f_{past}^r is obtained by feature extraction with a social encoder $\mathcal{E}_{social}^r(\cdot)$.

Spatiotemporal encoder $\mathcal{E}_1^r(\cdot)$ extracts features from past trajectory, while social interactions between agents are encoded by interactive encoder $\mathcal{E}_2^r(\cdot)$. Spatial location information is analyzed using convolution operations, and spatiotemporal features are extracted with the utilization of GRU units. Following the GCN concept, agents' locations are represented as nodes $\mathbf{v}_i^r$, and their social interactions serve as edges $\mathbf{e}_{ij}^r$ between nodes. The initialization process is applied in (3)(4). By utilizing a message-passing mechanism between edges and nodes [44], [45], we obtain social interaction embedded features ($\mathbf{v}^r = \mathbf{v}_i^{r,k+1}$) for the target agent after updating edge and node information over k iterations. Mathematically

$$\mathbf{st}^r = \mathcal{E}_1^r\left(\mathbf{X}, \mathcal{X}_{\mathcal{N}}^r\right) = \mathcal{F}_{GRU}^r\left(\mathcal{F}_{Conv}^r\left(\mathbf{X}, \mathcal{X}_{\mathcal{N}}^r\right)\right) \tag{2}$$

$$\mathbf{v}_i^{r,0} = \mathcal{F}_v^{r,0}\left(\left\{\mathbf{x}_i^{(t)}\right\}_{t=-T_{pr}}^{0}\right) \tag{3}$$

$$\mathbf{e}_{ij}^{r,0} = \mathcal{F}_e^{r,0}\left(\left[\mathbf{v}_i^{r,0}; \mathbf{v}_j^{r,0}\right]\right) \tag{4}$$

$$\mathbf{v}_i^{r,k+1} = \mathcal{F}_v^{r,k+1}\left(\sum_{n\in\mathcal{N}} \mathbf{e}_{i,n}^{r,k}\right) \tag{5}$$

$$\mathbf{e}_{ij}^{r,k+1} = \mathcal{F}_e^{r,k+1}\left(\left[\mathbf{v}_i^{r,k+1}; \mathbf{v}_j^{r,k+1}\right]\right) \tag{6}$$

where $\mathcal{F}_{GRU}^r(\cdot)$ is GRU and $\mathcal{F}_{Conv}^r(\cdot)$ is convolutional layer, $\mathcal{F}_v^r(\cdot), \mathcal{F}_e^r(\cdot)$ are MLPs.

Past feature $f_{past}^r = [st^r; v^r]$ is obtained by concatenating and $\hat{\mathbf{y}}^{T_f,r}$ is decoded by intent decoder $\mathcal{D}_{int}^r(\cdot)$, which is MLP structure. To optimize the feature learning architecture, we utilize an intentional loss function $\mathcal{L}_{int}$

$$\hat{\mathbf{y}}^{T_f,r} = \mathcal{D}_{int}^r\left(\left[st^r; v^r\right]\right) \tag{7}$$

$$\mathcal{L}_{int} = \left|\left|\hat{\mathbf{y}}^{T_f,r} - \mathbf{y}^{T_f,r}\right|\right|_2^2. \tag{8}$$

For the module, which is an instance-based framework, completing the initialization of historical scenes, past features f_{past}^r are stored as history features f_{his}^r. For the ith feature $f_{his,i}^r$ in the initial historical scenes, we use its decoded past starting position $\hat{\mathbf{x}}_i^{-T_{pr}+1}$ and intention $\hat{\mathbf{y}}_i^{T_f}$ to filter similar stored instances. For the *ith* and *jth* feature, if the decoded history experience has close past starting positions and future intentions, then this pair of features is redundant, and one should be deleted. Mathematically, if the margin between them is less than the threshold, they are redundant in the following cases:

$$\left|\left|\mathbf{x}_i^{-T_{pr}+1} - \mathbf{x}_j^{-T_{pr}+1}\right|\right|_2 \le \theta_{his}, \left|\left|\mathbf{y}_i^{T_f} - \mathbf{y}_j^{T_f}\right|\right|_2 \le \theta_{des} \tag{9}$$

where θ_{his} and θ_{des} are two initial thresholds for filtering.

D. Goal Intent Prediction Module

When planning and executing a trajectory, humans typically hold an intention, while concurrently being influenced by the current scene, which encompasses various elements like agents, obstacles, and the environment. These elements can alter the extent of inclination toward the goal intent, meaning that the final intention may not perfectly align with the initial intention. It is instead refined through interactions. Once the scope of the goal intent is defined, it undergoes clustering, and its center serves as the predicted intention.

The task of the goal intent prediction module is to estimate the target agent's intentions in a specified time step. To dynamically capture fine-tuning and simulate intention changes, the module mainly models the influence degree of the interaction and constrains the shift scope of intention for yielding more accurate predictions. Given that coarse intention is $\hat{\mathbf{y}}_{cint}$, where $\hat{\mathbf{p}}_{cint}$ is extent of inclination corresponding to coarse intention. In order to get effective information from historical experience, we need to extract a certain amount of historical feature information from historical scene units to guide the prediction.

By calculating the correlation and sorting the historical features, we can get the historical scene features $f_{add}^r = [f_{his_1}^r, f_{his_2}^r, f_{his_3}^r, \ldots, f_{his_M}^r]$ that have the M_{sim}^r (the size of M_{sim}^r which is related to the time scale of trajectory information) highest similarity $\mathbf{s}_{sml,j}^r$ with the trajectory features in current scene in the historical scene unit. The correlation $\mathbf{s}_{sml,j}^r$ is obtained by comparing the similarity between past features f_{past}^r and historical features $f_{his,j}^r$, which can be calculated as follows:

$$\mathbf{s}_{sml,j}^r = \mathcal{F}_{add}^r\left(f_{past}^r, f_{his,j}^r\right), \quad j = 1, 2, 3, \ldots, \left|M_{sim}^r\right| \tag{10}$$

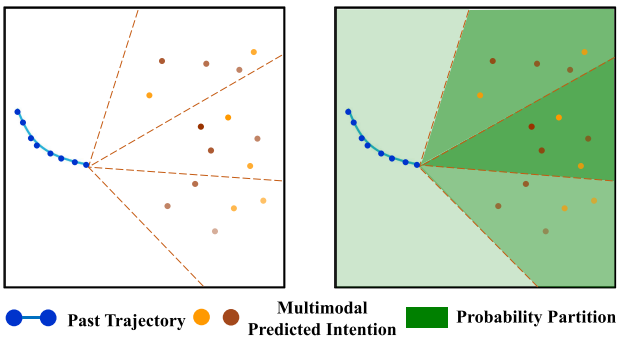

Fig. 3. Intention probability partition generation process: predictions for multiple scales are color-coded. The depth of the predictive points and partitions reflects the magnitude of probability.

where $\mathcal{F}^r_{add}(\cdot)$ consists of MLP structures with attention mechanisms for more appropriate difference measurement. coarse features f^r_{cint} are obtained by concatenating operation and then, f^r_{cint} is input into intent decoder $\mathcal{D}^r_{int}(\cdot)$ obtaining coarse intention $\hat{\mathbf{y}}^r_{cint} \in R^{M^r_{sim}}$. Mathematically

$$\hat{\mathbf{y}}^r_{cint} = \mathcal{D}^r_{int}\left(f^r_{cint}\right) = \mathcal{D}^r_{int}\left(\left[f^r_{past}; f^r_{add}\right]\right). \tag{11}$$

To avoid redundant intentions from affecting the subsequent clustering analysis, we drop out less relevant rough intentions by threshold thr^r_p. In addition, to mitigate duplication of multiscale information, we introduce weight coefficients based on spatial locations to impose constraints on the multiscale data

$$\hat{\mathbf{y}}_{cint}, \mathbf{s}_{sml} = \mathcal{F}_{sum}\left(\hat{\mathbf{y}}^r_{cint}, \mathbf{s}^r_{sml}\right) \tag{12}$$

where $\mathcal{F}_{sum}(\cdot)$ combines multiscale values with identical coordinates, including dropout operation.

Agents adapt their extent of inclination toward their goal intent and modify their trajectories when encountering obstacles en route to their destination. Uniform distribution alone would oversimplify this process and neglect nuanced changes in agents' intentions. To address this, we introduce the inclination extent parameter $\hat{\mathbf{p}}_{cint}$ to explicitly model scene interactions. Mathematically

$$\hat{\mathbf{p}}_{cint,i} = F_{softmax}\left(\mathbf{s}_{sml,i}\right) = \frac{e^{\mathbf{s}_{sml,i}}}{\sum_{l=1}^{M_{sim}} e^{\mathbf{s}_{sml,l}}} \tag{13}$$

where $\hat{\mathbf{p}}_{cint,i}$ and $\mathbf{s}_{sml,i}$ denote values for the ith historical feature, specifically referring to the extent of inclination and similarity score.

coarse intention $\hat{\mathbf{y}}^r_{cint}$ corresponds directly to the extent of inclination $\hat{\mathbf{p}}^r_{cint}$. To enhance the stability of intention location computation in the presence of motion noise, we employ softargmax as a more robust method for approximating the most accurate intention $\hat{\mathbf{y}}_{int,max}$. Mathematically

$$\hat{\mathbf{p}}_{int,max} = F_{softargmax}\left(\hat{\mathbf{p}}_{cint}\right) = \sum_{i=1} \frac{e^{\beta\hat{\mathbf{p}}_{cint,i}}}{\sum_{j=1} e^{\beta\hat{\mathbf{p}}_{cint,j}}} i. \tag{14}$$

For achieving reliable predictions, indiscriminate spatial sampling is suboptimal. Random sampling across regions may lead to inefficiency, potentially wasting samples in low-probability areas or repeatedly sampling neighboring regions of the same modality. This approach lacks the reliability required for accurate predictions. To obtain multimodal predicted intentions, we first estimate $\hat{\mathbf{p}}_{int,max}$ corresponding to the most probable coarse intention $\hat{\mathbf{y}}_{int,max}$, afterward clustering rough intentions $\hat{\mathbf{y}}_{cint}$. Mathematically

$$\hat{\mathbf{y}}_{supint} = \mathcal{F}_{kmeans}\left(\hat{\mathbf{y}}_{cint}\right) \tag{15}$$

$$\hat{\mathbf{p}}_{supint} = \mathcal{F}_{sumk}\left(\hat{\mathbf{p}}_{cint}\right). \tag{16}$$

After predicting possible trajectories, we employ the K-means clustering algorithm $\mathcal{F}_{kmeans}(\cdot)$ to generate $\hat{\mathbf{y}}_{supint} \in R^{K_e-1}$. $\mathcal{F}_{sumk}(\cdot)$ sums the extent of after cluster. After obtaining the clustering results, we generate multimodal predictions by concatenating $\hat{\mathbf{y}}_{int} = [\hat{\mathbf{y}}_{int,max}; \hat{\mathbf{y}}_{supint}], \hat{\mathbf{p}}_{int} = [\hat{\mathbf{p}}_{int,max}; \hat{\mathbf{p}}_{supint}]$, which includes the corresponding extent of inclination for each predicted intention. Unlike Y-Net [46], which relies on heatmap-based goal regression, our goal prediction module quantifies intent probability through instance-based historical matching. This avoids over-smoothing in sparse scenarios, offering improved performance in environments with limited trajectory data.

E. Dynamic Optimization Module

As additional information is acquired, the target agent steadily progresses toward their goal intent. During this progression, the likelihood of intentions aligning with goal intent direction increases, while the relative probability of other directions decreases.

To generate more accurate predicted trajectories for subsequent trajectory planning, along with dynamical decision making, in the dynamic optimization module, we introduce Bayesian optimization to optimize multimodal predicted intentions utilizing the motion time lapse, shown in Fig. 3. Equivalent to the combination of a priori and posteriori information, past observation and prediction results are combined with new results to constrain and modify the probability distribution.

Once goal intent prediction module obtains coarse intention $\hat{\mathbf{y}}^r_{cint}$ with $\hat{\mathbf{p}}^r_{cint}$. We perform multiscale reclustering of the coarse intentions and apply weighting factors to regulate the clustering outcomes, accounting for the varying degrees of influence from multiscale information on the future predictions. Subsequently, we employ kernel density estimation [47] to partition the clustering results into distinct spatial regions, where the value assigned to each region signifies the extent of agents' inclination toward reaching that specific region in the future. Mathematically

$$\hat{\mathbf{y}}^r_{kc}, \hat{\mathbf{s}}^r_{kc} = \mathcal{F}^r_{Kmeans}\left(\hat{\mathbf{y}}^r_{cint}, \hat{\mathbf{p}}^r_{cint}\right) \tag{17}$$

$$\hat{\mathbf{s}}^{r'}_{kc,j} = \hat{\mathbf{s}}^r_{kc,j} \times \omega^r \tag{18}$$

$$\hat{\mathbf{p}}_{pri,j} = \mathcal{F}_{Dy}\left(\mathbf{s}'_{kc,j}\right) = \frac{1}{Nh}\sum_{l=1}^{\sum K^r_c} \mathcal{F}_k\left(\frac{\mathbf{s}'_{kc,j} - \mathbf{s}'_{kc,l}}{h}\right) \tag{19}$$

$$\mathcal{F}_k(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{20}$$

where s_j and s_l are similarity scores of multiscale clustered intentions, the similarity scores for each time scale are multiplied by their respective weights. $\mathcal{F}_k(\cdot)$ is kernel function,

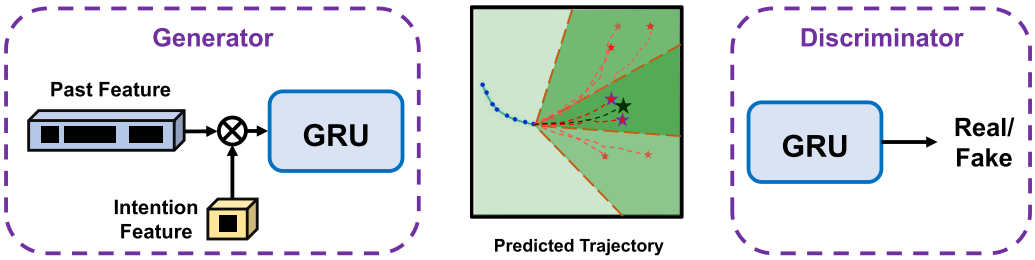

Fig. 4. Details of the trajectory generation module.

and h is smoothing parameter. The mentioned estimation is employed to derive inclination degree partitions, and individual points within each partition are assigned a regional inclination degree.

Time-fragmented information and the guidance of the current scene for future prediction are explicitly characterized by the extent of inclination, where a higher degree of inclination represents a higher target expectation of the future with the agent and a higher likelihood that the region contains the actual outcome. The priori knowledge of probabilistic partitioning is then utilized to guide the subsequently generated prediction intentions; prediction intention points $\hat{\mathbf{p}}_{int}$ obtained from the prediction model are multiplied by the degree of regional tendency of the regional distribution $\hat{\mathbf{p}}_{pri}$. Mathematically

$$\hat{\mathbf{p}}_{final} = \hat{\mathbf{p}}_{pri} \odot \hat{\mathbf{p}}_{int}. \tag{21}$$

Finally, the resulting $\hat{\mathbf{p}}_{final}$ is normalized to the top K_p ($K_p \leq K_e$) maximal values $\hat{\mathbf{p}}_{finali}{}^{max}_{K_p}$, which correspond to final prediction intentions $\hat{\mathbf{y}}_{finali}{}^{max}_{K_p}$.

F. Trajectory Generation Module

Based on the randomness of human motion, we use GAN [48] to fulfill the whole trajectory by getting the predicted intentions, shown in Fig. 4. The state of the target agent and the surrounding agents are input into DPIU as a condition to get the trajectory. By providing additional intent feature inputs to the generator and the discriminator, the network can be transformed into a conditional generative model (cGAN).

Inspired by [49], we employ an RNN-based generator due to the superior performance of RNNs in time series prediction problems. Assuming predicted intention $\hat{\mathbf{y}}_{int}$ of the agent's past trajectory $\mathbf{X}$ with its neighbor's past trajectories $\mathcal{X}_N$, the trajectory fulfilling process is

$$f_{traj} = \left[f_{past}^{full}; \mathcal{F}_d\left(\hat{y}_{final}\right)\right] \tag{22}$$

$$f_{htraj}^{t} = \mathcal{F}_{GRU}\left(f_{htraj}^{t-1}, f_{traj}\right) \tag{23}$$

$$\hat{\mathbf{y}}^{t} = \mathcal{F}_{Gtraj}\left(f_{htraj}^{t}\right) \tag{24}$$

where $\mathcal{F}_d(\cdot)$ is the MLP, trajectory features f_{traj} are obtained by concatenating past features f_{past}^{full} with encoded features $\mathcal{F}_d(\hat{y}_{final})$, f_{traj} are used as inputs to the GRU unit of the encoder at time and implemented RNN network recursion. To use features as conditions, we initialize the hidden state of the decoder and decode f_{htraj}^{t} by the generator $\mathcal{F}_{Gtraj}(\cdot)$, which consists of an encoder with a GRU unit.

Similarly, the discriminator consists of a GRU unit. Specifically, $Z_{real} = [\mathbf{X}, \mathbf{Y}]$ or $Z_{fake} = [\mathbf{X}, \hat{\mathbf{Y}}]$ are taken as inputs and categorized as real/fake. We add an MLP to the last hidden state of the encoder to obtain a classification score. Ideally, the discriminator learns the social interaction rules of pedestrians and classifies socially unacceptable trajectories as "Fake."

The adversarial loss $\mathcal{L}_{Adv}$ is the objective function of this module. In addition, we add $\mathcal{L}_{traj}$ to obtain the error value between the generated trajectory and the ground truth trajectory. Mathematically

$$\begin{aligned}\mathcal{L}_{Adv}\left(\mathcal{F}_G, \mathcal{F}_D\right) &= \min_G \max_D V\left(\mathcal{F}_G, \mathcal{F}_D\right) \\ &= E_{x\sim p(x)}\left[\log \mathcal{F}_G\left(x\right)\right] \\ &\quad + E_{z\sim p(z)}\left[\log\left(1 - \mathcal{F}_D\left(\mathcal{F}_G\left(z\right)\right)\right)\right]\end{aligned} \tag{25}$$

$$\mathcal{L}_{traj}\left(\hat{\mathbf{Y}}, \mathbf{Y}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left|\left|\hat{\mathbf{Y}}_i - \mathbf{Y}_i\right|\right|^2. \tag{26}$$

IV. Experiments

A. Datasets

SDD [50]: The dataset is a large dataset based on individual trajectories of traffic on the Stanford campus, while the information on these trajectories is captured through a bird's eye view. The dataset covers 20 on-campus scenarios and contains more than 11 000 different agent trajectories. It has been widely used in short-term trajectory prediction studies. As in most of the experiments, we sample the trajectories with FPS = 2.5 on the video information and generate past trajectories of length 8 (3.2 s) and predicted trajectories of length 12 (4.8 s).

ETH-UCY: The ETH/UCY benchmarks have been widely used for benchmarking trajectory forecasting models in the short-horizon setting in recent years. We used the pedestrian motion information provided in the original datasets of the ETH [51] scenes, ETH and HOTEL, and the UCY [52] scenes UNIV, ZARA1, and ZARA2. We analyzed a total of 1536 agents, each with their respective trajectories, which were converted to world coordinates (in meters) for consistency. Similar to the SDD dataset, the video was sampled at a frame rate of FPS = 2.5, with input (3.2 s) trajectories and predicted (4.8 s) future trajectories.

ApolloScape [53]: ApolloScape dataset is an open source dataset in the field of autonomous driving launched by Baidu, focusing on complex urban road scenes, providing large-scale, high-quality 3-D target trajectories (pedestrians, vehicles, bicycles, etc.) through multimodal on-board sensors

(e.g., cameras, LiDAR, and high-precision GPS) acquisition [53]. This database has a total of 53 and 50 min of training and testing sequences captured at 2 fps. Therefore, the official setting of observed and predicted horizons is 6(3 s) and 6(3 s) time steps, respectively.

B. Implementation Details

We implement our network in PyTorch, with feature dimensions defined as 128 for historical scene information and 64 for target historical scene units. On SDD, we set thresholds θ_{his} and θ_{des} to 1 for filtering the initial historical scene units. For full-length information, the number of coarse predictions L^{full} is 120; considering the possibility of overlap of intent points, we retain more coarse predictions for subsequent processing for time-fragmented information, thus L^{quar} is 180 and L^{half} is 240. The dropout thresholds $thr_p^r = 0.8 \times L^r$ for intent prediction at each time scale, and weighting values for different scales ω are set as $\omega^{full} = 0.8$, $\omega^{quar} = 0.7$, $\omega^{half} = 0.6$. The weights ζ as well as η are 0.8 and 0.75, $K_c^{full} = 20$, $K_c^{quar} = 25$, and $K_c^{half} = 30$.

On ETH-UCY, for global information, we filter the initial historical scenes with $\theta_{his}, \theta_{des} = 0.02, 0.02$ and L^{full} of 320. The parameters of the time-fragmented historical scenes are $L^{quar} = 400$, and $L^{half} = 480$. The dropout threshold for intent prediction for each time scale $thr_p^r = 0.8 \times L^r$, and weighting values for different scales ω and the weights ζ as well as η, are set as above. Multiscale K_c^r is also set to the same.

On ApolloScape, for global information, we filter the initial historical scenes with $\theta_{his}, \theta_{des} = 1, 1$ and L^{full} of 240. The parameters of the time-fragmented historical scenes are $L^{quar} = 300$, and $L^{half} = 360$. The dropout threshold for intent prediction for each time scale $thr_p^r = 0.8 \times L^r$, and weighting values for different scales ω and the weights ζ, as well as η, are set as above. Multiscale K_c^r is also set to the same.

Train our network with the following hyperparameter settings: the number of training batches is 32, and we use an initial learning rate of 10^{-3} for the feature learning framework, 10^{-4} for the memory addresser, and 10^{-3} for the trajectory implementation. All these modules are fine-tuned with a learning rate of 10^{-6}, β was set to 1, with an epoch count of 600. We trained the entire framework with the ADAM optimizer. All experiments are conducted on a single RTX 3090 GPU and an Intel Xeon E5-2680 v4 CPU.

C. Metrics

We use the established average displacement error (ADE) and final displacement error (FDE) as a judge of the model's prediction ability. ADE is the average difference in distance between the predicted whole trajectory and the ground truth trajectory, and FDE is the difference in distance between the predicted endpoint and the ground truth value. According to what was written in previous studies, in the case of multiple predictions, the smallest error value among all predicted values is taken as the final error.

D. Quantitative Results

Comparisons With State-of-the-Art Methods: This section compares the proposed DPIU with various baseline methods. Details of the baseline methods used are given below:

1) *Social-GAN [49]:* Based on GAN, modeling interactions between pedestrians through social pooling for multimodal trajectory prediction.
2) *Trajectron++ [54]:* A graph-based recurrent model integrating scene semantics and interactions for multimodal probabilistic prediction.
3) *SOPHIE [6]:* A model integrates scene physical constraints and a social attention mechanism, generating pedestrian trajectories conforming to scene rules through GAN.
4) *MANTRA [32]:* Combining a memory network and attention mechanism, storing historical trajectory patterns through the memory module to enhance long-time-dependent modeling.
5) *PECNet [56]:* An improved version of a variational autoencoder generates diverse and physically feasible predictive trajectories based on potential spatial trajectory sampling guided by endpoint estimation.
6) *MemoNet [31]:* Capturing long-term patterns of pedestrian movement using a learnable memory module and enhancing trajectory prediction reliability through memory retrieval.
7) *GTPPO [28]:* A trajectory prediction model fuses pedestrian obstacle avoidance experience through the graph attention mechanism, combining temporal attention with physically aware latent variables generated by pseudo-predictive machines for multimodal trajectory prediction.
8) *TrafficPredict [58]:* An LSTM-based trajectory prediction method with a hierarchical architecture.
9) *STAR [10]:* It proposed a novel transformer-based graph convolution to extract the spatiotemporal embedding.
10) *TPNet [59]:* An innovative two-stage approach for motion prediction: generating candidate motion proposals and refining these proposals by incorporating physical constraints.
11) *STGAT [11]:* An autoencoder-based trajectory prediction method models dynamic interactions between pedestrians through a spatiotemporal graph attention network, combining attention weights in both temporal and spatial dimensions.
12) *Social-STGCNN [60]:* Pedestrian interactions are modeled through a graph structure, resulting in significant improvements in prediction accuracy and inference speed.
13) *Transformer-TF [61]:* An improved version of the transformer utilizes a multilayer self-attention mechanism to fuse spatiotemporal contexts.
14) *TP-EGT [29]:* A collision-aware graph transformer to model complex traffic participant interactions, combining multitask learning to synchronously predict trajectories and interaction probabilities to alleviate the problem of over-smoothing of the graph structure.

TABLE I

MINADE/MINFDE RESULTS FOR TRAJECTORY PREDICTION ON SDD (K = 20) ARE REPORTED, WHERE LOWER VALUES INDICATE BETTER PERFORMANCE. BOLD FONT REPRESENTS THE BEST RESULT (ADE/FDE IN PIXELS). OUR METHOD PERFORMS IMPRESSIVELY, ACHIEVING BOTH THE LOWEST FDE VALUE AND THE LOWEST ADE VALUE

Metric	SOPHIE [6] CVPR'19	Trajectron++ [54] ECCV'20	MANTRA [32] CVPR'20	GTPPO [28] TNNLS'22	SHENet [33] NIPS'22	MemoNet [31] CVPR'22	Meta-IRLSOT [30] ESWA'24	NPSN CVPR'22	LED [55] CVPR'23	DPIU Ours
ADE	16.27	11.92	8.96	10.13	9.01	8.56	8.56	8.57	8.48	**8.29**
FDE	29.38	16.42	17.76	15.35	13.24	12.66	12.53	11.85	**11.66**	12.12

TABLE II

MINADE/MINFDE RESULTS FOR TRAJECTORY PREDICTION ON ETH-UCY (K = 20) ARE REPORTED, WHERE LOWER VALUES INDICATE BETTER PERFORMANCE. BOLD FONT DENOTES THE BEST RESULT (ADE/FDE IN METERS). OUR METHOD ACHIEVES THE SECOND-LOWEST AVERAGE FDE VALUE AND THE LOWEST AVERAGE ADE VALUE, DEMONSTRATING STRONG PREDICTIVE CAPABILITY

Models	Subsets	ETH ADE	ETH FDE	HOTEL ADE	HOTEL FDE	UNIV ADE	UNIV FDE	ZARA1 ADE	ZARA1 FDE	ZARA2 ADE	ZARA2 FDE	AVG ADE	AVG FDE
Social-GAN [49]	CVPR'18	0.87	1.62	0.67	1.37	0.76	1.52	0.35	0.68	0.42	0.84	0.61	1.21
STGAT [11]	ICCV'19	0.65	1.12	0.35	0.66	0.52	1.10	0.34	0.69	0.29	0.60	0.43	0.82
MANTRA [32]	CVPR'20	0.48	0.88	0.17	0.33	0.37	0.81	0.27	0.58	0.30	0.67	0.32	0.65
PECNet [56]	CVPR'20	0.54	0.87	0.18	0.24	0.35	0.60	0.22	0.39	0.17	0.30	0.29	0.48
TP-EGT [29]	ITS'24	0.39	0.60	0.10	0.16	0.23	0.43	0.18	0.31	0.14	0.22	0.20	0.34
SHENet [33]	NIPS'22	0.37	0.58	0.17	0.28	0.26	0.43	0.21	0.34	0.18	0.30	0.24	0.39
MemoNet [31]	CVPR'22	0.40	0.61	0.11	0.17	0.24	0.43	0.18	0.32	0.14	0.24	0.21	0.35
LED [55]	CVPR'23	0.39	0.58	0.11	0.17	0.26	0.43	0.18	0.26	**0.13**	0.22	0.21	0.33
DSBCL-PRM [21]	TASE'24	0.37	0.53	0.11	0.17	0.22	0.40	0.18	0.33	**0.13**	0.24	0.20	0.33
GSTGM [57]	IVC'24	0.29	0.54	0.13	0.29	**0.19**	0.39	0.17	**0.25**	0.14	0.21	**0.18**	0.32
E-V2-Net-SC [22]	CVPR'24	**0.26**	**0.40**	0.14	0.18	0.22	**0.35**	0.18	0.31	**0.13**	0.24	**0.18**	**0.31**
DPIU	Ours	0.36	0.58	**0.09**	**0.14**	0.21	0.41	**0.15**	0.30	0.14	**0.21**	**0.18**	0.32

15) *SHENet [33]:* A collision-aware graph transformer to model complex traffic participant interactions, combining multitask learning to synchronously predict trajectories and interaction probabilities to alleviate the problem of over-smoothing of the graph structure.
16) *Meta-IRLSOT [30]:* A pedestrian trajectory prediction framework based on meta-inverse reinforcement learning, which aims to solve the problems of insufficient correlation between scene information and trajectory and weak generalization ability of small samples of new scenes in traditional methods.
17) *LED [55]:* A stochastic trajectory prediction framework based on jump diffusion processes is proposed to address the insufficient real-time performance and inadequate multimodal capture in traditional diffusion models.
18) *STI-TP [62]:* A spatiotemporal interleaved multimodal trajectory prediction model, which addresses the inadequacy of existing methods in modeling dynamic interactions and motion multimodality through multiscale temporal coding, cross-attention spatiotemporal interactions, and multimodal decoders.
19) *GSTGM [57]:* A multipath pedestrian trajectory prediction framework based on GCN, spatiotemporal attention mechanism, and generative model aims to solve the problems of insufficient correlation between scene information and trajectories.
20) *DSBCL-PRM [21]:* A model uses dynamic subclass-balancing contrastive learning to balance head/tail samples and progressive refinement blocks to enhance motion patterns and predictions.
21) *E-V2-Net-SC [22]:* A novel angle-based social interaction representation, which enhances multimodal prediction accuracy by providing a more comprehensive representation of agent interactions within dynamic environments.

In order to evaluate the performance of our method, we conducted comparative experiments with the state-of-the-art methods on SDD. Table I lists the results of experiments for a K value of 20. In all the methods, the relative displacement of the pedestrians was entered into the coding module in order to make a fair comparison. As shown in the table, the ADE/FDE of DPIU is 3.25%/4.46% lower than MemoNet (8.56/12.66) and outperforms the graph-attention-based GTPPO (10.13/15.35) and meta-learning driven Meta-IRLSOT (8.57/12.53). Unlike closed box models, e.g., Social-GAN's random noise, PECNet's implicit endpoints, DPIU generates interpretable multimodal trajectories by probabilizing the Goal Intent space and Bayesian optimization. In complex and densely populated intersection scenarios, DPIU achieves ADE that is 43.79% lower than that of Trajectron++, which relies on a static graph structure. The performance improvement is attributed to DPIU's multiscale detail feature module, which dynamically adjusts interaction weights, effectively mitigating the over-smoothing issue commonly associated with the fixed graph attention mechanism employed by GTPPO. DPIU outperforms existing methods in terms of prediction accuracy, interpretability through explicit intent-driven and dynamic Bayesian optimization, and its generalization property without scenario fine-tuning is more attuned to the complexity of the open road.

Table II presents the experiments between the proposed DPIU and state-of-the-art methods on ETH-UCY. DPIU with SHENet (0.24/0.39) achieves the lowest ADE, and the

TABLE III

MINADE/MINFDE OF TRAJECTORY PREDICTION ON APOLLOSCAPE ($K = 20$) ARE REPORTED, WHERE LOWER VALUES ARE BETTER. THE BOLD FONT INDICATES THE BEST RESULT (ADE/FDE IN METERS). OUR METHOD DEMONSTRATES SUPERIOR PERFORMANCE BY ACHIEVING BOTH THE LOWEST FDE AND ADE VALUES

Metric	Social-GAN [49] CVPR'19	TrafficPredict [58] AAAI'19	STAR [10] CVPR'20	TPNet [59] CVPR'20	TP-EGT [29] ITS'24	STI-TP [62] ECE'24	DPIU Ours
ADE	1.33	7.18	0.95	0.77	0.70	0.64	**0.53**
FDE	2.45	11.12	1.80	1.48	1.28	1.04	**0.91**

TABLE IV

MINADE/MINFDE OF TRAJECTORY PREDICTION FOR DIFFERENT K VALUES ON SDD. LOWER IS BETTER. OUR METHOD ACHIEVES GENERATING MORE ACCURATE PREDICTION INTENTS WITH SMALLER K VALUES (ADE/FDE IN PIXELS)

. ADE%/FDE% clearly demonstrates the change in model performance when the number of candidate trajectories is reduced by K.

metric	K value	Trajectron++ [54]	PECNet [56]	GTPPO [28]	Meta-IRLSOT [30]	MemoNet [31]	DPIU
ADE(ADE%)	20	11.92	9.96	10.13	8.57	8.56	**8.29**
	14	15.03(26.09%)	12.30(23.49%)	12.53(23.68%)	9.78(14.12%)	9.75(13.90%)	**9.40(13.39%)**
	8	19.96(32.08%)	15.02(22.11%)	15.88(26.74%)	11.50(17.59%)	11.47(17.64%)	**10.91(16.06%)**
	2	28.66(43.59%)	21.57(43.61%)	23.11(45.53%)	16.96(47.48%)	16.85(46.90%)	**15.56(42.62%)**
FDE(FDE%)	20	16.42	15.88	15.35	12.53	12.66	**12.13**
	14	21.71(32.22%)	19.26(21.28%)	18.75(22.15%)	15.52(23.86%)	15.62(23.38%)	**14.94(23.17%)**
	8	29.31(35.01%)	25.97(34.84%)	23.86(27.25%)	19.98(28.74%)	20.13(28.87%)	**18.82(25.97%)**
	2	47.92(63.49%)	42.65(64.23%)	40.13(68.19%)	33.24(66.37%)	34.08(69.30%)	**29.94(59.09%)**

FDE is reduced by 0.06 m compared to it, outperforming LED (0.21/0.33) with diffusion module and TP-EGT (0.26/0.44) with collision-aware graph Transformer. Instance-based approaches such as MANTRA and Memonet are highly interpretable and can utilize historical information to provide guidance for the current scenario. While DPIU optimizes closed box memory retrieval using explicit probabilistic models and utilizes multiscale information to complement sparse features, it achieves better prediction accuracy and provides a decision basis for the planning module. TP-EGT relies on the collision-aware graph Transformer and the interaction probability prediction to show more stable obstacle avoidance ability and prediction accuracy in dense interaction scenarios, which may deviate from the physical constraints, though avoiding conflicts. Our proposed approach, DPIU, demonstrates significant advancements in trajectory prediction through the incorporation of explicit intent expression and multiscale detail extraction. By leveraging these enhancements, DPIU effectively captures latent information embedded within pedestrian trajectories, enabling a more nuanced understanding of movement dynamics.

We also conducted experiments with other methods on ApolloScape, see Table III. The results show that the comprehensive performance of DPIU is better than that of existing mainstream models. In complex interaction scenarios, the ADE of DPIU is reduced by 45.3% compared to TPnet, while its FDE is further optimized by 40.7% compared to TP-EGT. Thanks to the dynamic interaction perception unit proposed by DPIU, which is able to adaptively capture the pedestrian-vehicle implicit social constraints between pedestrians and vehicles. Although Social-GAN generates diverse trajectories through adversarial training, its social pooling module, which relies on fixed rules, has limited performance in dense traffic scenarios. Similarly, Transformer-TF utilizes global temporal attention to improve the robustness of long-range prediction, but is less effective in high-conflict regions such as intersections because it does not explicitly model the interaction topology. In terms of collaborative multi-intelligence reasoning, benefit from the multiscale detail capture mechanism, DPIU reduces ADE by 32.1% and 20.8%, respectively, compared to TP-EGT and STI-TP that use transformers.

TABLE V

INFERENCE TIME COMPARISON OF VARIOUS MODELS. THE BOLD FONT REPRESENTS THE BEST RESULT

Model	Inference time
STAR [10]	123.2
Trajectron++ [54]	29.1
AgentFormer [23]	123.2
Social-STGCNN [60]	25.1
GSTGM [57]	17.7
MemoNet [31]	7.7
DSBCL-PRM [21]	**6.0**
DPIU	6.5

To evaluate the effectiveness of the proposed dynamic optimization module, we conducted comparative experiments on SDD by varying the number of multimodal samples K. As shown in Table IV, our method achieved higher prediction accuracy than other models at the same K value. In addition, we introduce a new metric, ADE/FDE Decline Rate (ADE%/FDE%), which can be formulated as follows:

$$ADE\% = \left(\frac{ADE_{biggerK}}{ADE_{smallerK}} - 1\right) \times 100\% \tag{27}$$

$$FDE\% = \left(\frac{FDE_{biggerK}}{FDE_{smallerK}} - 1\right) \times 100\%. \tag{28}$$

Fig. 5. Multimodal prediction by DPIU on SDD with 20 predicted intents. DPIU can provide diverse and accurate intent prediction.

TABLE VI

ABLATION EXPERIMENTS FOR EACH COMPONENT IN DPIU ON SDD (PIXELS) AND ETH-UCY (METERS). EACH COMPONENT IS BENEFICIAL. THE BOND FONT REPRESENTS THE BEST RESULT

Multi-scale Detail	Level of Convergence	Coarse Intention Dropout	SDD ADE	SDD FDE	ETH-UCY AVG ADE	ETH-UCY AVG FDE
			8.56	12.66	0.21	0.35
✓			8.48	12.49	0.21	0.35
✓	✓		8.33	12.37	0.20	0.34
✓	✓	✓	**8.29**	**12.13**	**0.18**	**0.32**

It is defined as the ratio of the ADE/FDE for larger K values minus the ADE/FDE for smaller K values minus one. This metric aims to evaluate the model's performance at lower prediction outputs as the number of multimodal samples K decreases. For example, when K decreases from 20 to 14, the performance degradation rate ADE%/FDE% of our method (e.g., ADE increases by 13.39%) is significantly lower than MemoNet's 13.90% under the same conditions, indicating stronger robustness under low-modality conditions. This capability enables the generation of reliable trajectories even under minimal intent candidate conditions, aligning with the computational efficiency requirements of real-time autonomous systems while maintaining decision interpretability through explicit intent probability distributions. These characteristics are critical for deploying trajectory prediction modules in latency-sensitive autonomous driving applications.

Table V presents a detailed evaluation of inference speeds of various trajectory prediction models. We benchmarked several representative models under unified inference configuration, including STAR, Trajectron++, AgentFormer, Social-GCNN, GSTGM, MemoNet, and DPIU. All models were trained and evaluated on the UCY/ETH dataset to ensure consistency and comparability. Inference time is a critical metric for real-world deployment, as it directly affects models' ability to predict in a timely and efficient manner. Among evaluated models, GSTGM demonstrates outstanding computational efficiency, achieving an inference time of 17.7 ms. This low latency allows GSTGM to maintain accurate predictions while significantly reducing computational overhead, making it well-suited for real-time applications. DSBCL-PRM demonstrates the fastest inference speed, with prediction errors on the ETH-UCY dataset comparable to those of DPIU. DPIU achieves an inference time of 6.5 ms, which is attributed to its reliance solely on trajectory coordinate data, avoiding the computational complexity of image-based input processing. This streamlined approach substantially accelerates the prediction process, making DPIU particularly effective in time-sensitive scenarios. Despite the fundamental differences in their underlying architectures and operational principles, they achieve similar levels of prediction accuracy and computational efficiency, indicating that their future integration could yield improved performance.

E. Qualitative Results

1) Visualization of Multimodal Intentions: Fig. 5 shows multimodal intent prediction with DPIU. We see that with the help of intention probability distributions as well as clustering, DPIU can provide diverse and accurate intention predictions as well.

2) Advantages of Instance-Based Modeling: Fig. 6 compares predicted trajectories generated by our DPIU with previous approaches, MemoNet, and PECNet. It can be seen that similar scenarios stored in the memory bank can provide specific knowledge to guide the prediction of multimodal future intentions, and at the same time, the memory pointer can be localized to its corresponding scenario memory to better restore historical scenario conditions that are similar to the current scenario, and to guide the prediction of trajectories in the current scenario.

3) Visualization of the Dynamic Optimization Module: Fig. 7 shows the multimodal intent prediction with DPIU versus the prediction using the Bayesian prediction optimization algorithm. We can see that the algorithm captures the future trends of pedestrian trajectories better.

F. Model Analysis

1) Selection of Historical Fragmented Information: We experimentally validate trajectories with varying observation time lengths, specifically defined as the difference in the number of frames between the observation start and end points. This includes the evaluation of full-length trajectories with an observation time length of 8 frames. By observing Fig. 8, it is evident that the prediction accuracy exhibits significant changes at observation lengths of four and six frames. Beyond four frames, the prediction accuracy declines sharply as the

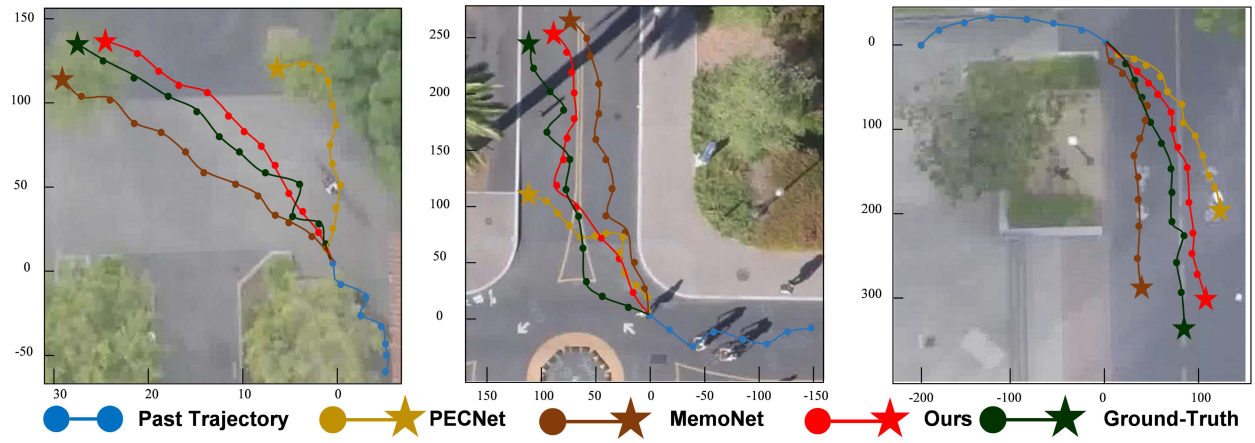

Fig. 6. Compare the best predicted trajectories produced by our method with the two previous similar methods on SDD. Our method achieves more accurate trajectory prediction.

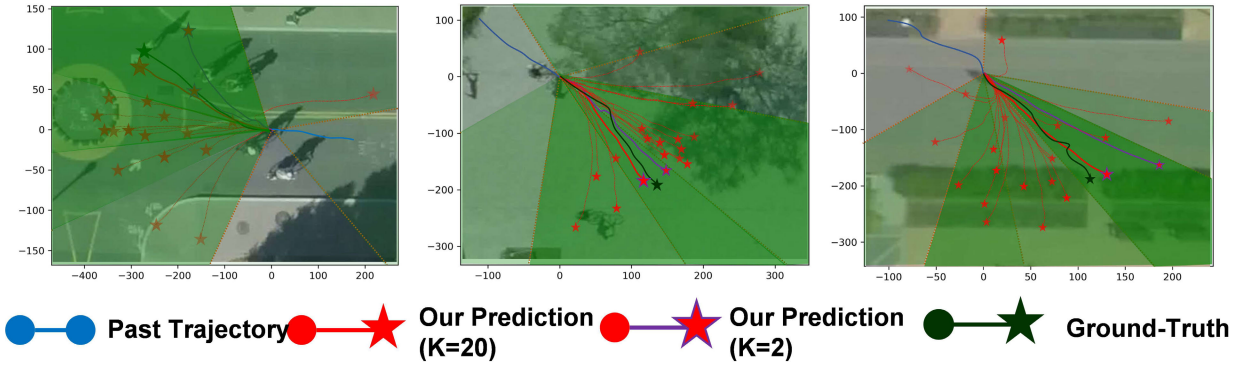

Fig. 7. Multimodal prediction versus exact prediction on SDD by DPIU.

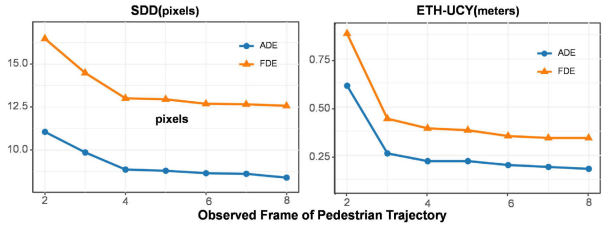

Fig. 8. Performance of different observation lengths in predicting the same future trajectory length on SDD/ETH-UCY.

number of frames decreases. Therefore, we select the past trajectories with observation lengths of four and six frames as the time fragment information.

2) Impact of the Number of Historical Scene Features: In order to obtain the appropriate number of historical scene features, we verified the effect of the number of historical scene feature information on the prediction performance in different time scales. Fig. 9 illustrates the effect of the number of historical scene features M. We find that when M is too large or too small, it will lead to the degradation of prediction performance, because: 1) when M is small, the model will miss some potentially valid historical experiences, leading to the lack of diversity and performance degradation; and 2) when M is too large, the current prediction will incorporate too many historical experiences with low relevance to the current one, which will lead to the deterioration of prediction performance.

3) Effect of Adjusted Time Windows: In order to verify the prediction ability of the model in dynamic traffic environments, we dynamically lengthen the observation time window and shorten the prediction time window, and verify the results. It can be seen in Fig. 9 that the prediction accuracy of our model increases with the increase of information obtained.

G. Ablation Experiment

We systematically evaluate our approach through a series of ablation experiments in Table VI, in which the following factors are examined: 1) remove the multiscale detail feature module and use only the entire history trajectory for prediction; 2) utilize all predictions equally without assigning probability values to coarse intentions; and 3) do not apply result dropout, and retain all prediction samples. Combining information from different dimensions effectively improves prediction performance. In addition, incorporating the extent of inclination into the model and applying partial dropout further enhances prediction accuracy.

By comparing whether the model includes multiscale information, we find that adding trajectory information from different dimensions expands the instance distribution,

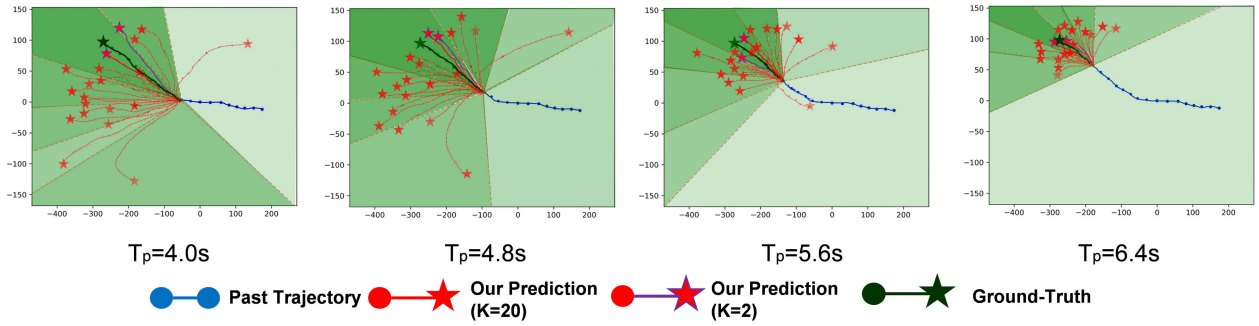

Fig. 9. Comparison of multimodal prediction versus accurate prediction by DPIU on SDD for different observation times.

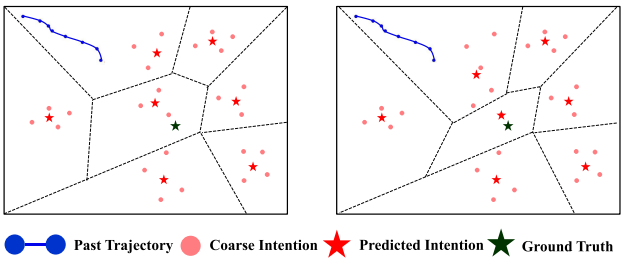

Fig. 10. Left: Noninclusion of pedestrian convergence on movement targets. Right: Inclusion of pedestrian convergence on movement targets.

enabling the model to acquire more useful information and thereby improving prediction performance.

Compared to other models, we introduce the degree of pedestrians' tendency toward the moving target. If the final prediction intention is obtained by only uniformly distributing the coarse intention, it will weaken the influence of those points that are more similar to the current situation on the prediction result, leading to a decrease in the prediction performance. By comparing Fig. 10, it can be seen that our model can better capture the variations in agent motion, which is closer to the motion patterns of agents in the real world and provides more accurate intention prediction.

Experimental observations demonstrate the substantial influence of various dropout values on prediction outcomes. Extreme dropout values affect diversity and degrade predictive performance. Overly large dropout values risk missing valuable predictive intentions; small dropout values can lead to the inclusion of numerous irrelevant instances, resulting in subsequent algorithmic complexities related to managing redundant data.

V. Conclusion

The work proposes DPIU, a novel trajectory prediction network that leverages the characteristics of human motion decision-making. It explicitly models goal intent selection by incorporating historical experience. The proposed framework systematically integrates three principal components: 1) a temporal-scale adaptive trajectory segmentation mechanism; 2) a probabilistic graphical model for hierarchical intention representation; and 3) a Bayesian inference-based optimization scheme for multimodal prediction generation. The proposed enhancements significantly improve the model's alignment with real-world traffic decision-making processes, thereby optimizing autonomous vehicle path-planning performance. Experimental results demonstrate state-of-the-art performance, achieving 5.2% lower ADE compared to baseline methods, which validates the critical role of incorporating human motion decision-making characteristics in trajectory prediction.

Limitation and Future Work: Analysis was limited to human motion decision-making traits and agent past trajectories, without modeling the associated traffic scenarios, resulting in a loss of environmental semantic information. Future research will incorporate traffic scenarios to influence agent behavior and employ scenario semantics for environmental condition predictions.

References

[1] X. Hu, L. Chen, B. Tang, D. Cao, and H. He, "Dynamic path planning for autonomous driving on various roads with avoidance of static and moving obstacles," *Mech. Syst. Signal Process.*, vol. 100, pp. 482–500, Feb. 2018.

[2] J. Yu and L. Petnga, "Space-based collision avoidance framework for autonomous vehicles," *Proc. Comput. Sci.*, vol. 140, pp. 37–45, Oct. 2018.

[3] R. Korbmacher and A. Tordeux, "Review of pedestrian trajectory prediction methods: Comparing deep learning and knowledge-based approaches," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24126–24144, Dec. 2022.

[4] H. Cheng, F. T. Johora, M. Sester, and J. P. Müller, "Trajectory modelling in shared spaces: Expert-based vs. Deep learning approach?," in *Proc. Int. Workshop Multi-Agent Syst. Agent-Based Simulation*, 2021, pp. 13–27.

[5] S. Zhang, M. Abdel-Aty, Q. Cai, P. Li, and J. Ugan, "Prediction of pedestrian-vehicle conflicts at signalized intersections based on long short-term memory neural network," *Accident Anal. Prevention*, vol. 148, Dec. 2020, Art. no. 105799.

[6] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1349–1358.

[7] H. Xue, D. Q. Huynh, and M. Reynolds, "PoPPL: Pedestrian trajectory prediction by LSTM with automatic route class clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 77–90, Jan. 2021.

[8] B. Yang, J. Zhu, Z. Yu, F. Fan, X. Liu, and R. Ni, "Fast adaptation trajectory prediction method based on online multisource transfer learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 1289–1304, 2025.

[9] R. Huang, H. Xue, M. Pagnucco, F. D. Salim, and Y. Song, "Vision-based multi-future trajectory prediction: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 8, pp. 13691–13708, Aug. 2025.

[10] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2020, pp. 507–523, doi: 10.1007/978-3-030-58610-2_30.

[11] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: Modeling spatial–temporal interactions for human trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6272–6281.

[12] P. Lv, W. Wang, Y. Wang, Y. Zhang, M. Xu, and C. Xu, "SSAGCN: Social soft attention graph convolution network for pedestrian trajectory prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 9, pp. 11989–12003, Sep. 2024.

[13] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, "Stepwise goal-driven networks for trajectory prediction," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2716–2723, Apr. 2022.

[14] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and U. Ozguner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 2, pp. 221–230, Jun. 2022.

[15] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, "Understanding and sharing intentions: The origins of cultural cognition," *Behav. Brain Sci.*, vol. 28, no. 5, pp. 675–691, Oct. 2005.

[16] A. D. Baddeley, *Essentials of Human Memory*. U.K.: Psychology Press, 2013.

[17] I. Ajzen and A. W. Kruglanski, "Reasoned action in the service of goal pursuit," *Psychol. Rev.*, vol. 126, no. 5, pp. 774–786, Oct. 2019.

[18] S. Chen, E. Yu, J. Li, and W. Tao, "Delving into the trajectory long-tail distribution for muti-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 19341–19351.

[19] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.

[20] L. F. Chiara, P. Coscia, S. Das, S. Calderara, R. Cucchiara, and L. Ballan, "Goal-driven self-attentive recurrent networks for trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2518–2527.

[21] B. Yang, K. Yan, C. Hu, H. Hu, Z. Yu, and R. Ni, "Dynamic subclass-balancing contrastive learning for long-tail pedestrian trajectory prediction with progressive refinement," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 8645–8658, 2025.

[22] C. Wong, B. Xia, Z. Zou, Y. Wang, and X. You, "SocialCircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 19005–19015.

[23] Y. Yuan, X. Weng, Y. Ou, and K. Kitani, "AgentFormer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9793–9803.

[24] C. Xu et al., "Dynamic-group-aware networks for multi-agent trajectory prediction with relational reasoning," *Neural Netw.*, vol. 170, pp. 564–577, Feb. 2024.

[25] C. Xu, M. Li, Z. Ni, Y. Zhang, and S. Chen, "GroupNet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6488–6497.

[26] J. Li, H. Ma, and M. Tomizuka, "Conditional generative neural system for probabilistic trajectory prediction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 6150–6156.

[27] J. Li, F. Yang, M. Tomizuka, and C. Choi, "EvolveGraph: Multi-agent trajectory prediction with dynamic relational reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 19783–19794.

[28] B. Yang, G. Yan, P. Wang, C.-Y. Chan, X. Song, and Y. Chen, "A novel graph-based trajectory predictor with pseudo-oracle," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7064–7078, Dec. 2022, doi: 10.1109/TNNLS.2021.3084143.

[29] B. Yang, F. Fan, R. Ni, H. Wang, A. Jafaripournimchahi, and H. Hu, "A multi-task learning network with a collision-aware graph transformer for traffic-agents trajectory prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6677–6690, Jul. 2024.

[30] B. Yang, Y. Lu, R. Wan, H. Hu, C. Yang, and R. Ni, "Meta-IRLSOT++: A meta-inverse reinforcement learning method for fast adaptation of trajectory prediction networks," *Expert Syst. Appl.*, vol. 240, Apr. 2024, Art. no. 122499.

[31] C. Xu, W. Mao, W. Zhang, and S. Chen, "Remember intentions: Retrospective-memory-based trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6478–6487.

[32] F. Marchetti, F. Becattini, L. Seidenari, and A. Del Bimbo, "MANTRA: Memory augmented networks for multiple trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7141–7150.

[33] M. Meng et al., "Forecasting human trajectory from scene history," in *Proc. Adv. Neural Inf. Process. Syst.*, Oct. 2022, pp. 24920–24933.

[34] P. S. Goldman-Rakic, "Cellular basis of working memory," *Neuron*, vol. 14, no. 3, pp. 477–485, Mar. 1995.

[35] A. D. Kiureghian and O. Ditlevsen, "Aleatory or epistemic? Does it matter?," *Struct. Saf.*, vol. 31, no. 2, pp. 105–112, Mar. 2009.

[36] D. Kahneman, *Thinking, Fast and Slow*. New York, NY, USA: Macmillan, 2011.

[37] D. Helbing, *Stochastische Methoden, Nichtlineare Dynamik Und Quantitative Modelle Sozialer Prozesse*. Aachen, Germany: Shaker, 1993.

[38] D. Wolinski, S. J. Guy, A. H. Olivier, M. C. Lin, D. Manocha, and J. Pettré, "Parameter estimation and comparative evaluation of crowd simulations," *Comput. Graph. Forum*, vol. 33, no. 2, pp. 303–312, 2014.

[39] A. J. Vingrys, "The visual brain in action," *Optometry Vis. Sci.*, vol. 74, no. 8, p. 596, Aug. 1997, doi: 10.1097/00006324-199708000-00016.

[40] R. T. Born and D. C. Bradley, "Structure and function of visual area MT," *Annu. Rev. Neurosci.*, vol. 28, no. 1, pp. 157–189, Jul. 2005, doi: 10.1146/annurev.neuro.26.041002.131052.

[41] K. Friston, "The free-energy principle: A unified brain theory?," *Nature Rev. Neurosci.*, vol. 11, no. 2, pp. 127–138, Feb. 2010, doi: 10.1038/nrn2787.

[42] R. B. Ivry and R. M. Spencer, "The neural representation of time," *Current Opinion Neurobiol.*, vol. 14, no. 2, pp. 225–232, Apr. 2004, doi: 10.1016/j.conb.2004.03.013.

[43] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach, "Intersubject synchronization of cortical activity during natural vision," *Science*, vol. 303, no. 5664, pp. 1634–1640, Mar. 2004, doi: 10.1126/science.1089506.

[44] J. Gilmer, S. S. Schoenholz, P. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1263–1272.

[45] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. S. Zemel, "Neural relational inference for interacting systems," in *Proc. Int. Conf. Mach. Learn.*, vol. 80, 2018, pp. 2688–2697.

[46] K. Mangalam, Y. An, H. Girase, and J. Malik, "From goals, waypoints & paths to long term human trajectory forecasting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15213–15222.

[47] R. A. Davis, K.-S. Lii, and D. N. Politis, "Remarks on some nonparametric estimates of a density function," in *Selected Works of Murray Rosenblatt* (Selected Works in Probability and Statistics), R. Davis, K. S. Lii, and D. Politis, Eds., New York, NY, USA: Springer, 2011, doi: 10.1007/978-1-4419-8339-8_13.

[48] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[49] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.

[50] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. ECCV*, 2016, pp. 549–565.

[51] S. Pellegrini, A. Ess, and L. V. Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *Proc. ECCV*, 2010, pp. 452–465.

[52] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Comput. Graph. Forum*, vol. 26, no. 3, pp. 655–664, 2007.

[53] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape open dataset for autonomous driving and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020, doi: 10.1109/TPAMI.2019.2926463.

[54] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: dynamically-feasible trajectory forecasting with heterogeneous data," in *Proc. ECCV*, 2020, pp. 683–700.

[55] W. Mao, C. Xu, Q. Zhu, S. Chen, and Y. Wang, "Leapfrog diffusion model for stochastic trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5517–5526.

[56] K. Mangalam et al., "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *Proc. ECCV*, 2020, pp. 759–776.

[57] M. H. K. Khel, P. Greaney, M. McAfee, S. Moffett, and K. Meehan, "GSTGM: Graph, spatial–temporal attention and generative based model for pedestrian multi-path prediction," *Image Vis. Comput.*, vol. 151, Nov. 2024, Art. no. 105245. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885624003500

[58] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha, "TrafficPredict: Trajectory prediction for heterogeneous traffic-agents," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Aug. 2019, pp. 6120–6127, doi: 10.1609/aaai.v33i01.33016120.

[59] L. Fang, Q. Jiang, J. Shi, and B. Zhou, "TPNet: Trajectory proposal network for motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6796–6805, doi: 10.1109/CVPR42600.2020.00683.

[60] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14412–14420.
[61] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 10335–10342.
[62] Y. Xu et al., "STI-TP: A spatio-temporal interleaved model for multi-modal trajectory prediction of heterogeneous traffic agents," *Comput. Electr. Eng.*, vol. 118, Aug. 2024, Art. no. 109361.

Jiaheng Xiao received the master's degree in traffic engineering from Jilin University, Changchun, China, in 2024.

He has worked on several funded projects in the context of connected vehicles and traffic safety. His research interests include pedestrian safety and computer vision applications.

Zhihui Li received the B.S. degree in computer science from East China University of Metallurgy, Ma'anshan, Anhui, China, in 2000, and the M.S. and Ph.D. degrees in computer software and theory from Jilin University, Changchun, China, in 2003 and 2007, respectively.

He is currently a Professor with the School of Transportation, Jilin University. His research interests include data mining, machine learning, signal processing, and intelligent traffic systems.

Mingxin Wang received the master's degree from the College of Communication Engineering, Jilin University, Changchun, China, in 2021. He is currently pursuing the Ph.D. degree with the Transportation College, Jilin University.

His research interests include digital image processing and pedestrian traffic characterization.

Yu Xie received the bachelor's degree in the Internet of Things Engineering from Chongqing Jiaotong University, Chongqing, China, in 2020, and the master's degree in electronic information from the Southwest University of Science and Technology, Mianyang, China, in 2023. He is currently pursuing the Ph.D. degree with the College of Transportation, Jilin University, Changchun, China.

His research interests include computer vision and pedestrian intention recognition.

Qin Ma received the master's degree from the Transportation College, Jilin University, Changchun, China, in 2024. He is currently pursuing the Ph.D. degree with the School of Navigation, Jimei University, Xiamen, China.

His research interests include vehicle formation strategies and vehicle-road collaboration.

Xin Wang received the master's degree from the Transportation College, Jilin University, Changchun, China, in 2024, where he is currently pursuing the Ph.D. degree.

His research interests include coupled sensing of the traffic environment and vehicle safety warning.

Yu Sun received the bachelor's degree in transportation engineering from Jilin University, Changchun, China, in 2021, where he is currently pursuing the Ph.D. degree in transportation engineering.

His main research interests include driver safety perception and early warning.