Understanding Interaction as You Need: Intention-Driven Pedestrian Behavior Prediction

Hang Yu[1], **Yansen Yu**[1], **Jiayan Qiu**[2*]

[1] School of Computer Engineering and Science, Shanghai University
[2] College of Computing and Mathematical Science, University of Leicester
{yuhang, yuys}@shu.edu.cn, jq46@leicester.ac.uk

Abstract

Prediction of pedestrian behavior is crucial for autonomous driving systems and intelligent transportation. Conventional methods predict the behavior based solely on either the pedestrian intention or the distance-related interactions between the pedestrian and its surroundings. However, these methods overlook the associations between intention and interaction for behavior prediction, in which they should be aligned with each other, thus leading to sub-optimal predictions. To solve this problem, we propose to predict the behavior by learning the association between intention and interaction, enabling them to mutually enhance each other during the prediction. Specifically, we first predict the short-term intention of all objects, including the target pedestrian and its surroundings. Then, instead of using the distance-related interactions, we predict the interactions by learning the correlated intentions. Finally, the intention-driven interactions refine the initial intention prediction, thus ensuring the alignment between intention and interaction for behavior prediction. We evaluate our method on two downstream tasks, the pedestrian trajectory prediction and pedestrian intention estimation, and show that it outperforms all the existing methods.

Introduction

Pedestrian behavior prediction, encompassing both pedestrian trajectory prediction (Jiang et al. 2025; Dong et al. 2023) and pedestrian intention estimation (Ham et al. 2023), is of pivotal importance for autonomous driving (Gu et al. 2023). In contrast to static and fully programmable objects, pedestrians change their short-term intention in a timely and dynamic manner by perceiving information from their surroundings and the environment (Yang et al. 2022). Therefore, the pedestrian behavior is highly nonlinear and time-varying.

Existing pedestrian behavior prediction approaches mainly focus on intention prediction and interaction modeling. The intention prediction methods aim to estimate the future goal or intention to cross the road of each pedestrian, without considering their interactions with the surroundings. Typically, the high-level intentions such as destinations (Mangalam et al. 2021, 2020) and road-crossing probabilities (Kotseruba, Rasouli, and Tsotsos 2021; Zhang, Tian, and Ding 2023) of the pedestrians are estimated from their motion history and scene context. The interaction modeling methods try to construct the interactions solely based on the visual distance, such as the distance between two pedestrians. Then, the interactions are modeled as dynamic graphs (Zong et al. 2024; Yao et al. 2024; Zhu et al. 2023), multihead attention (Yu et al. 2020; Wu et al. 2023; Yang et al. 2022), or social force maps (Ham et al. 2023; Yue, Manocha, and Wang 2022; Gupta et al. 2018), for predicting the pedestrian behavior.

However, as shown in Figure 1, intention-based approaches can identify plausible final goals but often miss the intermediate adjustments triggered by social interactions,

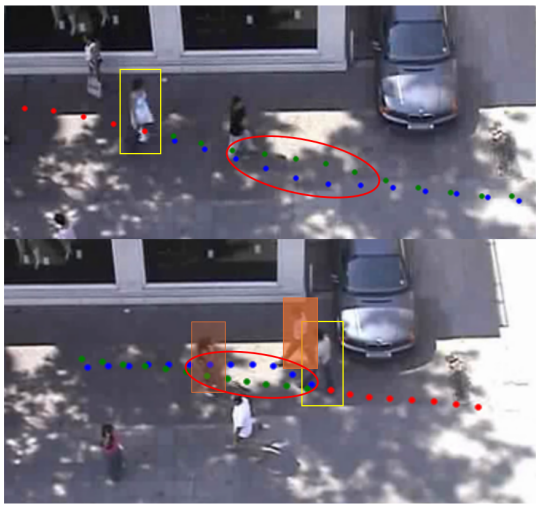

Figure 1: Illustration of limitations in existing pedestrian behavior prediction methods. (1) Top: Intention-based methods overlook the path adjustments of the pedestrian (marked by yellow box) caused by repulsion interaction with the one ahead, leading to biased intermediate prediction (red ellipse). (2) Bottom: Distance-based methods overemphasize the closer pedestrian (darker orange box) while underweighting distant one (lighter orange box), failing to predict the avoidance behavior (red ellipse).

*Corresponding author

producing deviations along the predicted path. On the other hand, distance-based assumptions could misjudge the most relevant agent to the future motion of the pedestrians by prioritizing the closest individuals, where, in fact, the agent that is far away poses a greater constraint. This misprioritization distorts the representation of the true avoidance strategy and ultimately degrades prediction quality.

To address the above limitations, we present an intention-driven interaction modeling framework that leverages intention to more effectively capture social interactions. Specifically, the framework first estimates each pedestrian's short-term intention using historical trajectory data and scene context. Then it explicitly detects potential intention conflicts among pedestrians to infer socially meaningful interactions that are likely to change future behaviors. These interactions are subsequently leveraged to iteratively refine the initial intentions, generating refined intentions that better reflect social interactions. Finally, the refined intentions guide downstream predictors, enabling behavior predictions that align with real-world pedestrian interaction dynamics.

We validate our framework across two representative tasks: pedestrian trajectory prediction and pedestrian crossing intention estimation. Experimental results demonstrate state-of-the-art performance across multiple key evaluation metrics, supplemented by intuitive visualizations that clearly illustrate how intention conflicts dynamically shape interactions and resulting behaviors. The main contributions of our work are summarized as follows:

- We redefine pedestrian interactions explicitly as conflicts in short-term intentions, establishing a novel, interpretable modeling paradigm.
- We introduce an intention-driven interaction modeling network that iteratively refines intentions based on interaction inference within a unified predictive framework.
- We achieve leading performance on multiple publicly available benchmarks, providing clear explanatory insights into the predicted pedestrian behaviors.

Related Work

Pedestrian behavior prediction aims to anticipate future actions in complex road scenes by fusing appearance, motion history, social cues, and context (Jiang et al. 2025; Bae, Lee, and Jeon 2024). Within this field, trajectory prediction seeks precise future paths from observed tracks, scene layout, and pedestrian interactions (Shi et al. 2023b; Feng et al. 2024), whereas intention estimation infers high-level goals such as crossing decisions from subtle behavioral and contextual cues (Chen et al. 2024; Ham et al. 2023). Recent work leverages transformers, graph structures, and multimodal fusion to move beyond reactive extrapolation toward cognitively and socially informed models (Wen, Yu, and Zheng 2025; Li, Yu, and Qiu 2025; Yu et al. 2024). Nevertheless, many methods embed interactions implicitly or overlook long-range intentions, hampering interpretability and accuracy (Wang et al. 2024; Shi et al. 2023a). These gaps (Xu et al. 2024) motivate our integrated framework, which explicitly models both intention and interaction to deliver reliable and socially compliant predictions.

Pedestrian trajectory prediction predicts future paths from past motion and scene context. Existing methods fall into two main lines. Interaction-centric models encode social or scene interactions via attention (Yu et al. 2020; Wu et al. 2023), graphs (Wong et al. 2024; Salzmann et al. 2020; Li, Ma, and Tomizuka 2019), or interpretable social-force formulations (Helbing and Molnar 1995; Yue, Manocha, and Wang 2022) that fit goal-driven and repulsive forces. Goal-driven models first sample a likely destination, then generate trajectories conditioned on that goal, using multi-scale goal inference to capture uncertainty (Wang et al. 2022; Mangalam et al. 2020, 2021). However, future-aware methods such as FRM (Park et al. 2023) model interaction only as a latent variable during decoding, so the alignment between agents' goals and social relations remains implicit. Interaction-based methods capture local social dynamics but tend to overlook longer-term intentions, while goal-based methods focus on final destinations yet struggle to model detailed interaction patterns, underscoring the need for approaches that can reason jointly about both goals and interactions (Xu et al. 2022).

Pedestrian intention estimation is mainly studied on JAAD (Kotseruba, Rasouli, and Tsotsos 2016) and PIE (Rasouli et al. 2019), which provide short driving clips with frame-level 2D pedestrian boxes and intention labels. Early methods (Razali, Mordan, and Alahi 2021) rely on single-frame CNNs or pose cues, while later works employ sequence models such as CNN–RNN hybrids (Chen, Tian, and Ding 2021), 3D CNNs (Singh and Suddamalla 2021; Yan et al. 2025), and graph convolutional networks (Scaccia, Pro, and Amerini 2025) to fuse appearance, pose, and coarse scene context. More recent approaches (Wang, Lai et al. 2024; Yang et al. 2022) introduce attention and transformer architectures to capture long-range spatio-temporal patterns and integrate multimodal signals (e.g., box trajectories, vehicle kinematics, semantic layouts). However, most of these models output only a single crossing probability, model interactions implicitly, and rarely establish an explicit link between a pedestrian's intention and the predicted motions of surrounding agents.

Preliminaries

Pedestrian Trajectory Prediction

Given the scene RGB image $I \in \mathbb{R}^{H\times W\times 3}$ and the observed trajectories of M pedestrians over past t_p time steps $\{X_i\}_{i=1}^{M}$, where $X_i = (o_1, o_2, \ldots, o_{t_p})$ and $o_t = (x_t, y_t) \in \mathbb{R}^2$ represents the 2D world coordinates of pedestrian i at time step t, trajectory prediction (Mangalam et al. 2021) aims to predict the future possible trajectories $\{\hat{Y}_i\}_{i=1}^{M}$, where $\hat{Y}_i = (o_{t_p+1}, \ldots, o_{t_p+t_f})$.

Pedestrian Intention Estimation

In the pedestrian intention estimation (Rasouli et al. 2019), the system receives an observation window of t_p ego-centric video frames and must decide whether the target pedestrian will step off the curb within the next t_f frames. For each time step $t \in [1, t_p]$ the raw input consists of the RGB

image $I_t \in \mathbb{R}^{H\times W\times 3}$, an axis-aligned 2D bounding box $b_t = (x_t^{\text{ul}}, y_t^{\text{ul}}, x_t^{\text{br}}, y_t^{\text{br}})$ that localizes the pedestrian, and an ego-vehicle state vector a_t describing the traffic context (instantaneous speed in the PIE dataset or discrete driver action in the JAAD dataset). We aggregate these observations into the spatiotemporal tensor $\mathcal{O} = \{(I_t, b_t, a_t)\}_{t=1}^{t_p}$. The prediction model then outputs a Bernoulli parameter $\hat{p} \in [0, 1]$ representing the probability that a crossing will begin in the interval $[t_p + 1, t_p + t_f]$; a binary decision $\hat{I} \in \{0, 1\}$ is obtained by thresholding $\hat{p}$ at a fixed value (typically 0.5).

Methods

We propose an interaction-centric framework to explicitly model social interactions in dynamic scenes and generate socially consistent predictions. As shown in Figure 2, the model first predicts each pedestrian's short-term intention using motion history and scene context, ignoring others to capture uninfluenced goals. An interaction-driven refinement module then estimates pairwise interactions from the initial intentions and iteratively refines intentions based on the estimated social relations. The refined intentions can better capture each pedestrian's anticipated behavior under both goal-directed drive and social interaction. Finally, a task-specific decoder maps the refined intentions to either a future trajectory prediction or a crossing-intention estimation.

Initial Intention Generation

To align the motion information with the scene information, we follow the trajectory processing method proposed by Y-Net (Mangalam et al. 2021), transform the trajectory into heatmaps $H_t \in \mathbb{R}^{t_p\times H\times W}$, defined as:

$$H(t, i, j) = 2\frac{\|(i, j) - o_t\|}{\underset{(x,y)\in I}{max}\|(x, y) - o_t\|}, \tag{1}$$

where (i, j) denotes pixel coordinates, and o_t represents the coordinates of the pedestrian at time step t. Meanwhile, the scene RGB image I is processed into a semantic segmentation map S (Qiu et al. 2021) using a pre-trained semantic segmentation network (Ronneberger, Fischer, and Brox 2015), where $H_s \in \mathbb{R}^{N_c\times H\times W}$, and N_c denotes the number of scene attributes such as walkable and non-walkable areas. Subsequently, the trajectory heatmap H_t is concatenated with the scene semantic map H_s along the channel dimension to generate the input tensor $H_i \in \mathbb{R}^{(t_p+N_c)\times H\times W}$. H_i is fed into the encoder network for further processing.

The encoder adopts a U-Net backbone (Ronneberger, Fischer, and Brox 2015). Specifically, we utilize a hierarchical architecture comprising multiple encoding blocks. At each encoding stage, convolutional and pooling operations progressively reduce spatial dimensions and simultaneously increase the number of feature channels, generating multi-scale feature representations denoted as $H_m = \{H_1, H_2, \ldots, H_6\}$. For three-dimensional inputs, these operations naturally extend to 3D convolutions and pooling, effectively capturing spatial correlations along all tensor dimensions. Mirroring the encoder structure, the decoder progressively restores the spatial resolution of feature maps via bilinear or trilinear upsampling operations, paired with convolutional refinement at each decoding stage. Skip connections fuse decoder activations H'_{i+1} with corresponding encoder features H_i:

$$H'_i = \text{Conv}\big([\mathcal{U}(H'_{i+1}),\, H_i]\big), \quad i = 5, 4, 3, 2, 1, \tag{2}$$

where $\mathcal{U}(\cdot)$ denotes the upsampling operation.

The final decoded feature map H_u passes through a pixel-wise sigmoid to produce a set of intention heatmaps, denoted collectively as $P(o_t, t)$, each encoding the probability that a pedestrian will occupy a given spatial location at a future time step. This unified representation serves as the initial intention and is forwarded to the subsequent interaction-driven refinement stage.

Interaction-Driven Intention Refinement

The intention probability distribution $P(o_t, t)$ represents the agent's future behavior trends. However, due to the influence of surroundings in the scene, an agent's future trajectory or crossing intention should arise from both goal-driven behavior and interaction dynamics. Therefore, explicitly modeling these interactions is crucial for refining agents' intention, enabling more realistic predictions for both trajectory prediction and intention estimation.

Unlike previous methods (Yue, Manocha, and Wang 2022) that directly derive interaction relationships from historical motion features, we argue that interactions are fundamentally intention-driven. Specifically, interactions emerge when the short-term intention of two agents, such as two pedestrians in the trajectory prediction task or a pedestrian and the ego-vehicle in the pedestrian intention estimation task, are correlated. These interactions iteratively refine the initial intention. At each future time step, we model interaction dynamics based on intention correlations between relevant agent pairs and use these interactions to iteratively refine their intention. This alternating process of interaction modeling and intention refinement progressively enhances prediction realism across all future prediction time steps.

Interaction modeling. For each time step $t \in [t_p+1, t_p+t_f]$, the future intention probability distribution of agent i, denoted as P_i, is first processed by a CNN to extract spatial features. An LSTM network then captures temporal dependencies by processing the sequential spatial features spanning from $t_p + 1$ to $t_p + t_f$:

$$H_i^{\text{st}} = \text{LSTM}\big(\{\text{Conv}(P_i(t))\}_{t=t_p+1}^{t_p+t_f}\big), \tag{3}$$

where $H_i^{\text{st}} \in \mathbb{R}^{t_f\times C}$ represents the spatio-temporal features of agent i across future time steps.

In the trajectory prediction task, the interaction feature between pedestrian i and another pedestrian j is computed using their respective spatio-temporal features. For the pedestrian intention estimation task, the interaction feature is computed between the pedestrian and the ego-vehicle. For each agent pair (i, j), the interaction vector is calculated as:

$$Z_{\text{interact}} = \text{MLP}(H_i^{\text{st}}, H_j^{\text{st}}), \tag{4}$$

where Z_{interact} is a time-varying vector representing the interaction strength between agents i and j.

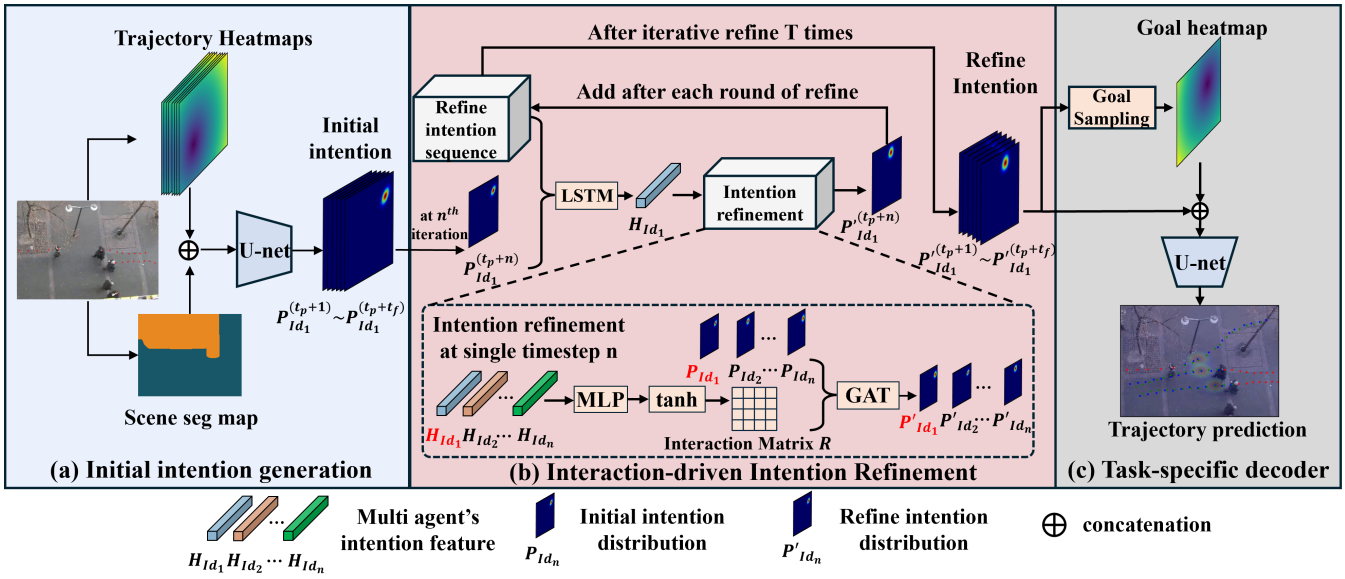

Figure 2: Overall architecture of the proposed framework. First, historical motion and scene information are used to generate an initial set of pedestrian intention estimates over t_f prediction steps. Next, at each future step, pairwise interactions are inferred from these intentions and used to iteratively refine intentions, where the refined intention sequence is initially empty and updated with each refinement. Finally, the refined intentions are fed into a task-specific decoder to produce the outputs for downstream tasks.

Then we obtain a sequence of continuous interaction coefficients $R_{ij}(t) = \tanh(Z_{\text{interact}}(t))$ in the range $(-1, 1)$, and interpret negative values as repulsion, values near zero as approximately neutral, and positive values as attraction.

Intention refinement. For each neighbor j of agent i, we use its intention map $P_j(t)$ and the time-dependent relation map $R_{ij}(t)$, both defined for $t = 1, \ldots, t_f$, to compute an overall score between agents i and j. Specifically, we first aggregate the time-varying features into a scalar

$$e_{ij} = \text{Pool}\big(P_j(t) \oplus R_{ij}(t)\big), \tag{5}$$

where Pool denotes global average pooling over both temporal and spatial dimensions, summarizing the interaction between agents i and j over the prediction horizon into a single scalar. We then normalize these scores over all neighbors of agent i to obtain attention weights:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in N(i)} \exp(e_{ik})}, \tag{6}$$

where the scalar α_{ij} measures the overall influence of neighbor j on agent i.

Given these weights, the refined intention map for agent i is computed as

$$P_i' = \text{MLP}\Big(P_i \oplus \sum_{j \in N(i)} \alpha_{ij} \cdot (P_j \oplus R_{ij})\Big), \tag{7}$$

where $\oplus$ denotes channel-wise concatenation.

By replacing P_i with the refined P_i' and repeating this process for T times, we iteratively incorporate interaction feedback from surrounding agents and obtain socially more plausible intention maps, which in turn lead to more accurate trajectory predictions and intention estimation.

Task-Specific Decoder

In the task-specific decoding stage, the model maps each refined intention to the required output format. Although we design separate decoders for trajectory prediction and intention estimation, they share the common processing pipeline introduced above.

Pedestrian trajectory prediction. After obtaining the refined intention probability distribution of the pedestrian, we use a trajectory generator to generate K possible future trajectories for the pedestrian.

First, as pedestrian trajectories are inherently goal-driven, we sample K possible goal points to effectively guide trajectory generation. Using the Test-Time Sampling Trick (TTST) from Y-Net (Mangalam et al. 2021), the pedestrian's goal probability distribution $P_{t_p+t_f}(x, y)$ is processed to determine the most probable goal point G_a via soft-argmax. A large number of candidate points are then sampled from $P_{t_p+t_f}(x, y)$, filtering out those with probabilities below $0.01 \times \max(P_{t_p+t_f})$. The remaining points are subsequently clustered into $K-1$ groups, with the cluster centers selected as goal points. Combined with G_a, these collectively form the goal set G containing K goal points.

After obtaining G, we convert these K goal points into heatmap representations and concatenate them with the first $(t_f - 1)$ refined intent maps $P_i'(x, y, t)$ (excluding the final goal time step). The concatenated features are then fed into a U-Net network to generate K sets of accurate future probability distribution $P_n(x, y, t)$.

Finally, we extract deterministic trajectory coordinates by applying a spatial soft-argmax to the normalized intention maps. For each time step t, we first obtain a normalized in-

tention distribution

$$\tilde{P}_n(x, y, t) = \frac{\exp(P_n(x, y, t))}{\sum_{x',y'} \exp(P_n(x', y', t))}, \tag{8}$$

and then compute the expected position via

$$(x_t, y_t) = \sum_{x,y} (x, y)\, \tilde{P}_n(x, y, t). \tag{9}$$

Pedestrian intention estimation. To effectively leverage refined pedestrian intention heatmaps for improved crossing intention estimation, we introduce a novel feature fusion strategy based on the base model proposed in TrEP (Zhang, Tian, and Ding 2023). Specifically, after intention refinement, we obtain a sequence of refined 3D pedestrian intention heatmaps across time steps. These heatmaps inherently capture the probabilistic spatial-temporal distribution of pedestrians' future locations and intentions.

We first employ a 3D convolutional neural network (3D CNN) to extract spatial-temporal features from these refined intention heatmaps, producing a compact intention representation $F_{\text{intention}}$. In parallel, the base model from TrEP encodes historical 3D trajectories and vehicle motion information through Transformer layers, generating feature representations F_{base}.

Subsequently, we fuse these two complementary features via an attention-based multilayer perceptron, where the attentionMLP adaptively weights and integrates relevant aspects from both sources. Specifically, we compute the fused feature as:

$$F_{\text{fusion}} = \text{MLP}\left(\text{attn}(F_{\text{intention}}, F_{\text{base}}) \,\|\, F_{\text{base}}\right), \tag{10}$$

where $\text{attn}(\cdot)$ denotes a single-head additive attention mechanism that selects relevant cues from the intention representation $F_{\text{intention}}$ using F_{base} as context, and $\|$ denotes feature concatenation. The resulting F_{fusion} is then passed to a softmax classifier for final intention prediction.

Experiment

Experimental Setup

Datasets. To evaluate our method across both trajectory prediction and intention estimation tasks, we conduct experiments on four widely-used datasets:
(1) **ETH** (Pellegrini et al. 2009) and **UCY** (Lerner, Chrysanthou, and Lischinski 2007) datasets are classic benchmarks for pedestrian trajectory prediction. They contain 1536 annotated trajectories across 5 subsets and 4 real-world scenes. The trajectories are labeled in meters, and we follow the leave-one-out evaluation protocol and adopt the preprocessing strategy from Trajectron++ (Salzmann et al. 2020).
(2) **Stanford Drone Dataset (SDD)** (Robicquet et al. 2016) contains over 11,000 pedestrian trajectories across 8 scenes with top-down drone views. Compared to ETH/UCY, SDD features more diverse and dense human interactions in complex open-space environments.
(3) **PIE** (Rasouli et al. 2019) and **JAAD** (Kotseruba, Rasouli, and Tsotsos 2016) datasets are used for pedestrian intention estimation. These datasets consist of video sequences captured from an ego-vehicle perspective, with annotations indicating whether pedestrians intend to cross the street. We reconstruct the 3D trajectories of both pedestrians and the ego-vehicle using COLMAP, based on the 15-frame historical segments provided in the datasets. This 3D reconstruction enables more fine-grained reasoning about interactions and spatial configurations, especially important for modeling short-term human intention.

Metrics. We evaluate the performance of our model on two tasks: pedestrian trajectory prediction and pedestrian intention estimation, each with task-specific metrics.

For pedestrian trajectory prediction, we adopt two standard metrics: Average Displacement Error (ADE) and Final Displacement Error (FDE) (Yue, Manocha, and Wang 2022). ADE is the mean Euclidean distance between the predicted and ground truth positions over all predicted time steps. FDE computes the Euclidean distance between the predicted final position and the ground truth final position.

For pedestrian intention estimation, we evaluate the model using four commonly used metrics: Accuracy, AUC, F1-score, and Precision (Kotseruba, Rasouli, and Tsotsos 2021). Accuracy denotes the proportion of correctly predicted crossing intentions. Area Under the ROC Curve (AUC) measures the model's ability to distinguish between crossing and non-crossing classes. F1-score is the harmonic mean of precision and recall, offering a balanced measure under class imbalance. Precision quantifies the ratio of true positive predictions among all positive predictions, reflecting the reliability of crossing predictions.

Implementation Details. This section summarizes the key training settings employed in our experiments.

The network is trained with three objectives: intention losses $L_{\text{intention}}$ and L_{refine} as binary cross-entropy between heatmaps and Gaussian-rendered positions; a trajectory loss L_{traj} using ℓ_2 regression on coordinates; and an intention-estimation loss L_{cross} as binary cross-entropy on the predicted crossing probability, combined into a loss L_{total}. And the number of iterations T is set to 2.

For pedestrian trajectory prediction, we employ the ADAM optimizer with a fixed learning rate of 1×10^{-4}, a batch size of 8, and train the model for 300 epochs. To increase data diversity, each scene image and its trajectory annotations are augmented by rotation and flipping at 90° intervals, resulting in an eight-fold expansion of training samples. To stabilize interaction modeling and accelerate convergence, we update only the intention generator during the first 100 epochs, and fine-tune the full network thereafter.

For pedestrian intention estimation, to obtain geometrically consistent pedestrian–vehicle interactions, we augment the raw annotations with a 3D reconstruction stage. For each video sequence, we apply COLMAP (Schonberger and Frahm 2016) to recover camera poses $\{T_t^{cv}\}_{t=1}^{t_p}$ in a common world frame. The ego-vehicle trajectory $v_t \in \mathbb{R}^3$ is extracted from camera centers, while pedestrian locations $p_t \in \mathbb{R}^3$ are estimated by back-projecting the bottom-center pixel of each 2D bounding box onto the ground plane. Collect-

Methods	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG	SDD
S-GAN (Gupta et al. 2018)	0.81/1.52	0.72/1.61	0.60/1.26	0.34/0.69	0.42/0.84	0.58/1.18	27.23/41.44
CGNS (Li, Ma, and Tomizuka 2019)	0.62/1.40	0.70/0.93	0.48/1.22	0.32/0.59	0.35/0.71	0.49/0.97	15.60/28.20
NEXT (Liang et al. 2019)	0.73/1.65	0.30/0.59	0.60/1.27	0.38/0.81	0.31/0.68	0.46/1.00	-
STAR (Yu et al. 2020)	0.36/0.65	0.17/0.36	0.31/0.62	0.26/0.55	0.22/0.46	0.26/0.53	-
PECNet (Mangalam et al. 2020)	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48	9.96/15.88
Trajectron++ (Salzmann et al. 2020)	0.39/0.83	0.12/0.21	0.20/0.44	**0.15**/0.33	**0.11**/0.25	0.19/0.41	-
Y-Net (Mangalam et al. 2021)	0.28/0.33	0.10/0.14	0.24/0.41	0.17/0.27	0.13/0.22	0.18/0.27	7.85/11.85
MSRL (Wu et al. 2023)	0.28/0.47	0.14/0.22	0.24/0.43	0.17/0.30	0.14/0.23	0.19/0.33	8.22/13.39
PPT (Lin et al. 2024)	0.36/0.51	0.11/0.15	0.22/0.40	0.17/0.30	0.12/0.21	0.20/0.31	7.03/10.65
DSTIGCN (Chen et al. 2025)	0.43/0.70	0.22/0.41	0.25/0.45	0.20/0.37	0.17/0.32	0.25/0.45	-
NSP-SFM (Yue, Manocha, and Wang 2022)	0.25/**0.24**	**0.09/0.13**	0.21/0.38	0.16/0.27	0.12/**0.20**	0.17/0.24	6.52/10.61
SocialCircle (Wong et al. 2024)	0.25/0.38	0.12/0.14	0.20/**0.34**	0.18/0.29	0.13/0.22	0.17/0.27	6.54/**10.36**
Ours	**0.24/0.24**	**0.09**/0.14	**0.19/0.34**	**0.15/0.25**	0.12/**0.20**	**0.16/0.23**	**6.49**/10.57

Table 1: Performance comparison on ETH & UCY and SDD datasets. Prediction uses 20 samples, and the minimum error is reported. For ETH/UCY, the unit is meters, and for SDD is pixels. Lower is better.

Methods	PIE				JAAD			
	Accuracy	AUC	F1	Precision	Accuracy	AUC	F1	Precision
MM-LSTM (Aliakbarian et al. 2018)	0.85	0.85	0.76	0.70	0.81	0.78	0.60	0.53
PCPA (Kotseruba, Rasouli, and Tsotsos 2021)	0.87	0.85	0.78	0.74	0.84	0.78	0.60	0.54
BiPed (Rasouli, Rohani, and Luo 2021)	0.92	0.91	0.86	0.83	0.84	0.80	0.62	0.54
TrEP (Zhang, Tian, and Ding 2023)	0.94	**0.95**	0.88	0.90	0.92	0.87	0.71	0.72
Ours	**0.95**	**0.95**	**0.90**	**0.93**	**0.93**	**0.89**	**0.73**	**0.74**

Table 2: Performance comparison on the PIE and JAAD datasets in the 3D setting. Higher values indicate better performance.

ing $\{(p_t, v_t)\}_{t=1}^{t_p}$ yields temporally aligned 3D trajectories for pedestrians and vehicles in the same coordinate frame, enabling accurate estimation of distance, relative velocity, and time-to-collision. These sequences are used to construct a geometry-aware observation tensor $\tilde{\mathcal{O}} = \{(p_t, v_t)\}_{t=1}^{t_p}$, which serves as input to the intention generation network. We train this task using ADAM with a learning rate of 5×10^{-3}, batch size of 8, and a total of 500 training epochs. To ensure fair comparison, all baseline methods are provided with the same 3D input sequences.

Experimental Results

Result for Pedestrian trajectory prediction. Table 1 presents quantitative comparisons on the ETH, UCY, and SDD datasets. All methods generate 20 samples per prediction and report the minimum ADE/FDE, following standard evaluation protocols. The comparison includes approaches that model pedestrian interactions using social force models, such as NSP-SFM (Yue, Manocha, and Wang 2022), as well as goal-driven trajectory prediction methods, such as Y-Net (Mangalam et al. 2021).

Our model achieves best performance across most ETH and UCY scenarios, with the lowest overall ADE/FDE averages of $0.16/0.23$. In particular, it performs best in four out of five scenes, including in the high-density UNIV scenario, which poses higher demands on interaction modeling, where our approach yields noticeable gains (ADE/FDE: $0.19/0.34$) over prior methods.

On the challenging SDD dataset, our method obtains an ADE of 6.49 and FDE of 10.57, marginally outperforming the previous best SocialCircle (Wong et al. 2024) and NSP-SFM (Yue, Manocha, and Wang 2022). Although FDE is not the absolute lowest, our consistent superiority across datasets and metrics confirms the effectiveness and generalizability of our intention-driven interaction modeling.

Result for Pedestrian intention estimation. Table 2 presents results on the PIE and JAAD benchmarks under the 3D evaluation setting. Our model consistently outperforms recent state-of-the-art methods, including the transformer-based TrEP (Zhang, Tian, and Ding 2023), across all reported metrics. It is worth noting that we do not compare with methods using pose or skeletal inputs, as these introduce additional modalities beyond our trajectory–scene–based framework.

These gains can be attributed to the explicit integration of short-term intention reasoning and interaction modeling within our framework. By leveraging 3D spatial cues and refining intention representations through interaction feedback, our model offers a more faithful interpretation of pedestrian decision-making.

Ablation Study

To assess the effectiveness of the proposed interaction modeling module, we conduct ablation studies on both trajectory prediction and intention estimation tasks, as shown in Table 3 and Table 4. Removing the interaction module yields a

Methods	ETH & UCY	SDD
	ADE / FDE	ADE / FDE
w/o interaction	0.24 / 0.35	8.12 / 12.39
w/ interaction	**0.16 / 0.23**	**6.49 / 10.57**

Table 3: Ablation study on pedestrian trajectory prediction.

Methods	PIE				JAAD			
	Acc	AUC	F1	Prec	Acc	AUC	F1	Prec
w/o interaction	0.91	0.89	0.86	0.83	0.84	0.80	0.62	0.53
w/ interaction	**0.95**	**0.95**	**0.90**	**0.93**	**0.93**	**0.89**	**0.73**	**0.74**

Table 4: Ablation study on pedestrian intention estimation.

variant that relies solely on individual intention cues without modeling social interactions.

Experimental results show consistent performance drops across both tasks when the interaction module is disabled. Specifically, the performance drop is more pronounced on the SDD and JAAD datasets, where social interactions are more complex. These results validate the effectiveness of our proposed intention-driven interaction modeling module. By explicitly modeling how individual intentions interact with surroundings, our framework better captures socially compliant behaviors and improves prediction quality in both spatial and decision-level dimensions.

Case Study

To further verify the effectiveness of our proposed solution in modeling complex interactions, we selected cases for analysis on the pedestrian trajectory prediction and pedestrian intention estimation tasks.

Figure 3 illustrates the effectiveness of our intention-driven interaction modeling for pedestrian trajectory prediction. The top row compares predictions before (left) and after (right) intention refinement. At T+0, pedestrians A and B have overlapping short-term intention distributions, indicating a strong mutual repulsive relation. Consequently, B allocates more attention to the distant A than to the closer C. After refinement, both A and B adjust their predicted paths to reduce collision risk, aligning more closely with the ground truth. At $T+3$, although B is close to both A and C, their intentions do not conflict, so attention to both remains low and the predictions stay accurate. At $T+4$, the interaction between A and B has vanished; B maintains only weak attention to the nearby C, and the prediction remains consistent with the ground truth. This case demonstrates that our approach can dynamically adjust interaction modeling based on evolving intentions, yielding socially plausible and accurate trajectory forecasts.

In the left part of Figure 4, yellow bounding boxes mark the pedestrian's past positions, the green box indicates the current position, and the red arrow shows high attention to the vehicle. In the right part, blue points and red points denote the vehicle's observed and predicted trajectories, while yellow and red triangles denote the pedestrian's observed

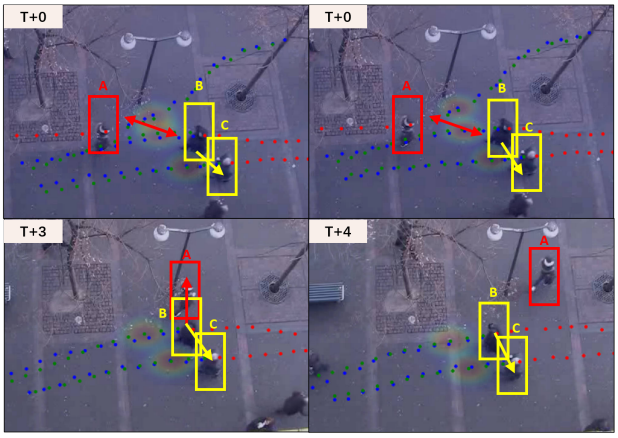

Figure 3: Accurate estimation of interactions based on intention improves trajectory prediction accuracy.

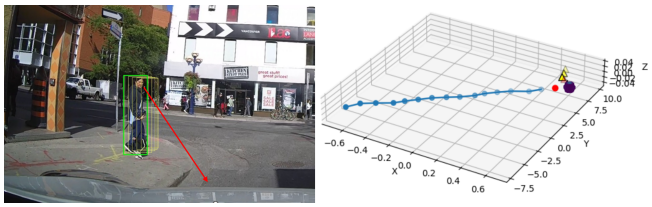

Figure 4: Illustration of correctly capturing interaction of vehicle and pedestrian in a high-uncertainty scenario. When the pedestrian and the ego vehicle are in an ambiguous negotiation state, the model successfully estimates that the pedestrian's behavior in the upcoming frames will be crossing.

and predicted trajectories. The heatmap reveals strong short-term interaction between the pedestrian and vehicle, suggesting a negotiation state. Historical trajectories and refined intentions indicate that the vehicle decelerates in the final moments, reflecting a proactive yielding strategy. Consequently, the model infers a higher crossing probability. This case demonstrates that intention-based interaction modeling provides crucial cues for accurate crossing-intention estimation in high-uncertainty scenarios.

Conclusion

We propose an intention-driven interaction modeling framework for pedestrian behavior prediction. By explicitly treating interactions as conflicts among short-term intentions, the framework enables interpretable reasoning about social dynamics. Through iterative refinement between intention and interaction, it captures causal effects that are often overlooked by prior methods. Extensive experiments across benchmarks demonstrate the effectiveness and generality of our approach, validating its potential as a principled foundation for future socially-aware prediction systems.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62302287), and by projects of the Shanghai Committee of Science and Technology, China (Grant No. 23ZR1423500).

References

Aliakbarian, M. S.; Saleh, F. S.; Salzmann, M.; Fernando, B.; Petersson, L.; and Andersson, L. 2018. VIENA: A driving anticipation dataset. In *Asian Conference on Computer Vision*, 449–466. Springer.

Bae, I.; Lee, J.; and Jeon, H.-G. 2024. Can Language Beat Numerical Regression? Language-Based Multimodal Trajectory Prediction. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 753–766.

Chen, T.; Tian, R.; and Ding, Z. 2021. Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3103–3109.

Chen, W.; Sang, H.; Wang, J.; and Zhao, Z. 2025. DSTIGCN: Deformable Spatial-Temporal Interaction Graph Convolution Network for Pedestrian Trajectory Prediction. *IEEE Transactions on Intelligent Transportation Systems*.

Chen, X.; Zhang, S.; Li, J.; and Yang, J. 2024. Pedestrian crossing intention prediction based on cross-modal transformer and uncertainty-aware multi-task learning for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 25(9): 12538–12549.

Dong, Y.; Wang, L.; Zhou, S.; and Hua, G. 2023. Sparse instance conditioned multimodal trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9763–9772.

Feng, Y.; Carballo, A.; Fujii, K.; Karlsson, R.; Ding, M.; and Takeda, K. 2024. MulCPred: Learning Multi-Modal Concepts for Explainable Pedestrian Action Prediction. *Sensors*, 24(20): 6742.

Gu, J.; Hu, C.; Zhang, T.; Chen, X.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5496–5506.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2255–2264.

Ham, J.-S.; Kim, D. H.; Jung, N.; and Moon, J. 2023. Cipf: Crossing intention prediction network based on feature fusion modules for improving pedestrian safety. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3666–3675.

Helbing, D.; and Molnar, P. 1995. Social force model for pedestrian dynamics. *Physical review E*, 51(5): 4282.

Jiang, J.; Yan, K.; Xia, X.; and Yang, B. 2025. A Survey of Deep Learning-Based Pedestrian Trajectory Prediction: Challenges and Solutions. *Sensors (Basel, Switzerland)*, 25(3): 957.

Kotseruba, I.; Rasouli, A.; and Tsotsos, J. K. 2016. Joint attention in autonomous driving (JAAD). *arXiv preprint arXiv:1609.04741*.

Kotseruba, I.; Rasouli, A.; and Tsotsos, J. K. 2021. Benchmark for evaluating pedestrian action prediction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1258–1268.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. In *Computer graphics forum*, volume 26, 655–664. Wiley Online Library.

Li, J.; Ma, H.; and Tomizuka, M. 2019. Conditional generative neural system for probabilistic trajectory prediction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6150–6156. IEEE.

Li, R.; Yu, H.; and Qiu, J. 2025. Dynamic Shadow Unveils Invisible Semantics for Video Outpainting. In *NeurIPS*.

Liang, J.; Jiang, L.; Niebles, J. C.; Hauptmann, A. G.; and Fei-Fei, L. 2019. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5725–5734.

Lin, X.; Liang, T.; Lai, J.; and Hu, J.-F. 2024. Progressive pretext task learning for human trajectory prediction. In *European Conference on Computer Vision*, 197–214. Springer.

Mangalam, K.; An, Y.; Girase, H.; and Malik, J. 2021. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15233–15242.

Mangalam, K.; Girase, H.; Agarwal, S.; Lee, K.-H.; Adeli, E.; Malik, J.; and Gaidon, A. 2020. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European conference on computer vision*, 759–776. Springer.

Park, D.; Ryu, H.; Yang, Y.; Cho, J.; Kim, J.; and Yoon, K.-J. 2023. Leveraging future relationship reasoning for vehicle trajectory prediction. *arXiv preprint arXiv:2305.14715*.

Pellegrini, S.; Ess, A.; Schindler, K.; and Van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*, 261–268. IEEE.

Qiu, J.; Yang, Y.; Wang, X.; and Tao, D. 2021. Scene Essence. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8318–8329.

Rasouli, A.; Kotseruba, I.; Kunic, T.; and Tsotsos, J. 2019. PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6261–6270.

Rasouli, A.; Rohani, M.; and Luo, J. 2021. Bifold and semantic reasoning for pedestrian behavior prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15600–15610.

Razali, H.; Mordan, T.; and Alahi, A. 2021. Pedestrian intention prediction: A convolutional bottom-up multi-task approach. *Transportation research part C: emerging technologies*, 130: 103259.

Robicquet, A.; Sadeghian, A.; Alahi, A.; and Savarese, S. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, 549–565. Springer.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Salzmann, T.; Ivanovic, B.; Chakravarty, P.; and Pavone, M. 2020. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, 683–700. Springer.

Scaccia, S.; Pro, F.; and Amerini, I. 2025. Unsupervised pedestrian intention estimation through deep neural embeddings and spatio-temporal graph convolutional networks. *Pattern Analysis and Applications*, 28(2): 108.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Shi, L.; Wang, L.; Long, C.; Zhou, S.; Tang, W.; Zheng, N.; and Hua, G. 2023a. Representing multimodal behaviors with mean location for pedestrian trajectory prediction. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 11184–11202.

Shi, L.; Wang, L.; Zhou, S.; and Hua, G. 2023b. Trajectory unified transformer for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9675–9684.

Singh, A.; and Suddamalla, U. 2021. Multi-input fusion for practical pedestrian intention prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2304–2311.

Wang, C.; Wang, Y.; Xu, M.; and Crandall, D. J. 2022. Stepwise goal-driven networks for trajectory prediction. *IEEE Robotics and Automation Letters*, 7(2): 2716–2723.

Wang, K.-L.; Tsao, L.-W.; Wu, J.-C.; Shuai, H.-H.; and Cheng, W.-H. 2024. TrajFine: Predicted trajectory refinement for pedestrian trajectory forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4483–4492.

Wang, T.-W.; Lai, S.-H.; et al. 2024. Multi-Modal Pedestrian Crossing Intention Prediction with Transformer-Based Model. *APSIPA Transactions on Signal and Information Processing*, 13(5).

Wen, J.; Yu, H.; and Zheng, Z. 2025. WeatherPrompt: Multimodality Representation Learning for All-Weather Drone Visual Geo-Localization. In *NeurIPS*.

Wong, C.; Xia, B.; Zou, Z.; Wang, Y.; and You, X. 2024. Socialcircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19005–19015.

Wu, Y.; Wang, L.; Zhou, S.; Duan, J.; Hua, G.; and Tang, W. 2023. Multi-stream representation learning for pedestrian trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2875–2882.

Xu, X.; Qiu, J.; Wang, X.; and Wang, Z. 2022. Relationship spatialization for depth estimation. In *European Conference on Computer Vision*, 615–637. Springer.

Xu, X.; Qiu, J.; Yu, B.; and Wang, Z. 2024. Visual Relationship Transformation. In *European Conference on Computer Vision*, 251–272. Springer.

Yan, Y.; Zhou, M.; Feng, C.-c.; Lv, L.; and Ding, H. 2025. Three-dimensional CNN-based model for fine-grained pedestrian crossing behavior recognition in automated vehicles. *Journal of Transportation Engineering, Part A: Systems*, 151(2): 04024116.

Yang, D.; Zhang, H.; Yurtsever, E.; Redmill, K. A.; and Özgüner, Ü. 2022. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles*, 7(2): 221–230.

Yao, P.; Zhu, Y.; Bi, H.; Mao, T.; and Wang, Z. 2024. TrajCLIP: Pedestrian trajectory prediction method using contrastive learning and idempotent networks. *Advances in Neural Information Processing Systems*, 37: 77023–77037.

Yu, C.; Ma, X.; Ren, J.; Zhao, H.; and Yi, S. 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European conference on computer vision*, 507–523. Springer.

Yu, H.; Li, R.; Xie, S.; and Qiu, J. 2024. Shadow-Enlightened Image Outpainting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7850–7860.

Yue, J.; Manocha, D.; and Wang, H. 2022. Human trajectory prediction via neural social physics. In *European conference on computer vision*, 376–394. Springer.

Zhang, Z.; Tian, R.; and Ding, Z. 2023. Trep: Transformer-based evidential prediction for pedestrian intention with uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3534–3542.

Zhu, W.; Liu, Y.; Wang, P.; Zhang, M.; Wang, T.; and Yi, Y. 2023. Tri-HGNN: Learning triple policies fused hierarchical graph neural networks for pedestrian trajectory prediction. *Pattern Recognition*, 143: 109772.

Zong, M.; Chang, Y.; Dang, Y.; and Wang, K. 2024. Pedestrian Trajectory Prediction in Crowded Environments Using Social Attention Graph Neural Networks. *Applied Sciences*, 14(20): 9349.