Intention-Aware Diffusion Model for Pedestrian Trajectory Prediction

Yu Liu[1,2], Zhijie Liu[1], Xiao Ren[1], Youfu Li[2*], He Kong[1*]

[1]Southern University of Science and Technology
[2]City University of Hong Kong
yuliu254-c@my.cityu.edu.hk,{12332642,12431359}@mail.sustech.edu.cn,
meyfli@cityu.edu.hk, kongh@sustech.edu.cn

Abstract

Predicting pedestrian motion trajectories is critical for the path planning and motion control of autonomous vehicles. Recent diffusion-based models have shown promising results in capturing the inherent stochasticity of pedestrian behavior for trajectory prediction. However, the absence of explicit semantic modelling of pedestrian intent in many diffusion-based methods may result in misinterpreted behaviors and reduced prediction accuracy. To address the above challenges, we propose a diffusion-based pedestrian trajectory prediction framework that incorporates both short-term and long-term motion intentions. Short-term intent is modelled using a residual polar representation, which decouples direction and magnitude to capture fine-grained local motion patterns. Long-term intent is estimated through a learnable, token-based endpoint predictor that generates multiple candidate goals with associated probabilities, enabling multimodal and context-aware intention modelling. Furthermore, we enhance the diffusion process by incorporating adaptive guidance and a residual noise predictor that dynamically refines denoising accuracy. The proposed framework is evaluated on the widely used ETH, UCY, NBA, and SDD benchmarks, demonstrating competitive results against state-of-the-art methods.

Code — https://github.com/AISLAB-sustech/IAD

Introduction

Pedestrian motion prediction is a critical capability for extensive applications, including autonomous driving (Li et al. 2021; Yang et al. 2024a,b), robots navigation (Eiffert et al. 2020), and planning (Yin et al. 2025). Given the observed trajectories of pedestrians, accurately forecasting their future paths is essential for ensuring safe and efficient operation. A major challenge lies in the inherently stochastic and non-deterministic nature of human motion, shaped by social interactions and environmental context. As pedestrian behavior unfolds over time, such influences manifest as both transient adjustments and global planning objectives. Addressing this challenge requires modelling both long-term destination goals and short-term motion intentions(Lin et al. 2024; Duan et al. 2022).

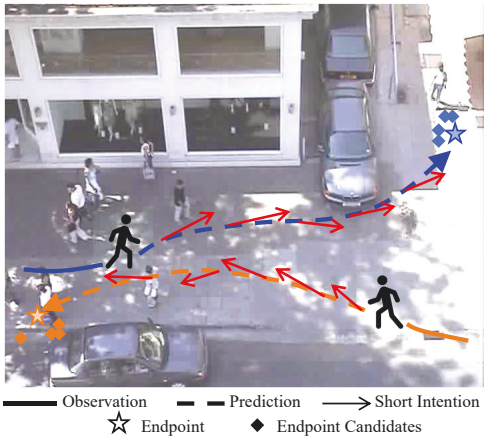

Figure 1: An illustration of pedestrian intentions. The polar-based short-term intentions and long-term multimodal endpoints provide motion cues that guide the generation of future trajectories via a diffusion process.

Within the above context, works such as (Zhao et al. 2021; Chiara et al. 2022; Xing et al. 2025) adopt goal-based approaches, where trajectory endpoints are first predicted before inferring intermediate positions. This strategy aims to capture long-term dependencies and reveal the global motion tendencies. On the other hand, some methods incorporate midway points (Bae and Jeon 2023; Bae, Park, and Jeon 2024; Kothari, Sifringer, and Alahi 2021), which explicitly model the trajectory's intermediate state to capture local dynamic patterns arising from fine-grained variations between consecutive timesteps. However, such anchor-based methods often rely on fixed spatial references to represent future motion, which may fail to capture the underlying semantics of pedestrian intent. These representations may misinterpret subtle variations in motion intention that are not truly indicative of a shift in the pedestrian's intended direction. For instance, a pedestrian slightly curving toward a building entrance may still maintain a steady forward-moving intent, whereas intermediate anchors might misleadingly indicate an unintended deviation or turning behavior.

Another challenge lies in modelling pedestrian motion intentions. Some trajectory forecasting methods (Wu et al.

*Corresponding authors.

2024; Liao et al. 2025) decompose intention into a set of discrete semantic categories, such as turning left, accelerating, or stopping. While this approach provides interpretability, it imposes rigid constraints on the inherently continuous and fine-grained nature of human movement. For example, a slight veer of 10 degrees and a sharp 90-degree turn may both be classified as turning left, despite their drastically different implications for trajectory evolution. Such categorical abstractions may oversimplify motion dynamics, making it difficult to capture the subtle variations and uncertainty that characterize real pedestrian behavior. These limitations highlight the need for a unified framework capable of capturing both the fine-grained nature of short-term motion and the uncertainty of long-term goals.

To tackle the above issues, our intention-aware design provides early and structured cues that are directly useful for downstream autonomous-driving planners, especially in safety-critical scenarios such as cut-ins or pedestrian crossings. In this paper, we propose an Intention-Aware Diffusion model (IAD) for pedestrian trajectory prediction, which captures both short-term motion tendencies and long-term goal hypotheses as illustrated in Figure 1. In the short-term intention module, a residual polar representation is employed to model pedestrian motion intent. By separately encoding direction and magnitude, this formulation offers a compact, continuous, and expressive representation capable of capturing fine-grained motion patterns. The residual design promotes a structured modelling paradigm, where global motion cues provide a coarse prior that is refined through local adjustments, reflecting the hierarchical nature of human navigation. For long-term intention estimation, we propose a learnable token-based estimator that predicts multiple candidate endpoints with associated probabilities, capturing uncertainty and intent diversity. The token encodes trajectory-level context, enabling goal hypotheses that are both multimodal and context-aware, while the probabilistic formulation supports more informed and interpretable predictions. In the diffusion model, a residual noise prediction module is introduced to estimate the discrepancy between predicted and true noise, enabling dynamic refinement of the denoising process for improved trajectory generation. Further, an auxiliary guidance strategy enables smooth integration of conditional signals and balances diversity and fidelity.

Related Works

Pedestrian Trajectory Prediction

Early methods formulate predicting pedestrians' trajectories based on traditional ways, including social force (Helbing and Molnar 1995), Kalman Filters (Klingelschmitt et al. 2014), and Markov models (Firl et al. 2012). However, they may struggle to represent the complexity and variability of pedestrian motion in dynamic and crowded situations. Modern learning-based approaches have achieved significant advances. These methods model trajectory forecasting as a sequential estimation task using deep sequential processing models. For instance, Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN) (Alahi et al. 2016) have been employed to capture pedestrian motion dynamics. Inspired by Natural Language Processing (NLP), Transformer models (Shi et al. 2023; Lin et al. 2024) have been introduced to better capture long-range dependencies and global context, leading to improved performance in trajectory prediction tasks. Given the importance of social interactions, many approaches use graph-based (Liu et al. 2023; Bae and Jeon 2023; Kim et al. 2024) structures to explicitly model complex inter-pedestrian relations.

Diffusion Models for Trajectory Prediction

Diffusion model (Sohl-Dickstein et al. 2015) is first proposed to solve non-equilibrium thermodynamics problems and have shown prior generation capabilities on various tasks, including image synthesis (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021), video generation (Ho et al. 2022), and natural language processing (Li et al. 2022). In the context of motion prediction, recent works have integrated diffusion models into prediction frameworks. MID (Gu et al. 2022) explicitly simulates the process of human motion variation from indeterminate to determinate via the reverse process of diffusion. LED (Mao et al. 2023) introduces a trainable leapfrog initialiser to skip denoising steps for prediction efficiency. TRACE (Rempe et al. 2023) constrains trajectories using target waypoints, speed, and specified social groups, while incorporating the surrounding context. DICE (Choi et al. 2024) introduces an efficient sampling mechanism coupled with a scoring module to select the most plausible trajectories. C2F-TP (Wang et al. 2025) proposes a coarse-to-fine prediction framework in which a conditional denoising model is used to refine the uncertainty of samples. However, these diffusion-based approaches do not consider motion intention in frameworks.

Prior Conditioned Approach

Due to the stochastic nature of human motion, pedestrians' trajectories contain randomness. To enable controllable and interpretable trajectory predictions, some approaches incorporate anchor points as prior information to guide the generation of multimodal trajectories. TNT (Zhao et al. 2021) formulates trajectory prediction as a two-stage process by first predicting discrete future endpoints and then generating target-conditioned trajectories. DenseTNT (Gu, Sun, and Zhao 2021) builds on this by replacing discrete endpoint classification with continuous density regression, improving spatial coverage and prediction accuracy.(Wang et al. 2023) generates diverse proposals fused with goal-oriented anchors to enable multimodal prediction. Graph-TERN (Bae and Jeon 2023) predicts intermediate control points by segmenting the future path, and refines trajectories using a spatiotemporal multi-relational graph for better accuracy. PPT (Lin et al. 2024) progressively trains the model through next-step prediction, destination prediction, and full trajectory prediction, capturing both short- and long-term motion patterns. SingularTrajectory (Bae, Park, and Jeon 2024) proposes an adaptive method which corrects misplaced anchors based on a traversability map. However, lacking semantic intent modelling may hinder prediction quality.

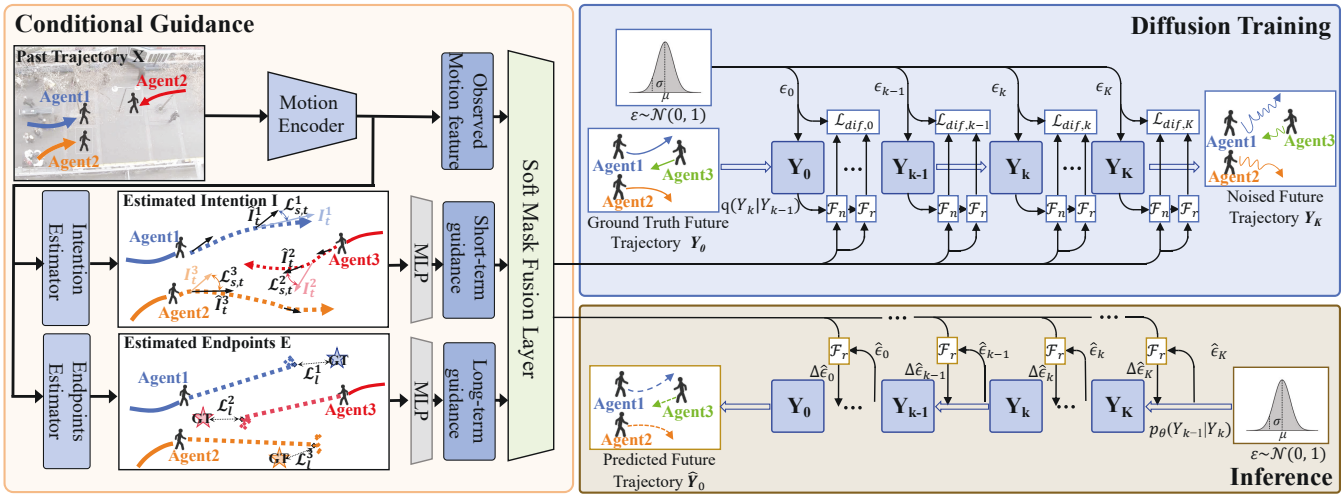

Figure 2: The overall architecture of the proposed framework. The conditional guidance module estimates both short-term and long-term motion intentions, which are integrated through a soft-mask fusion layer. In the diffusion module, the noise estimator $\mathcal{F}_n$ works in tandem with the refinement network $\mathcal{F}_r$ to iteratively denoise and generate future trajectories.

Methods

Preliminaries

The diffusion model comprises a forward process and a reverse process. In the forward process, Gaussian noise is incrementally added to a sample drawn from the data distribution $\{Y_k\}_{k=0}^{K} \in \mathbb{R}^{t_{pred}\times 2}$, which corresponds to the future trajectories. This process starts from ground truth path Y_0 and is repeated for K steps following a predefined noise schedule, gradually transforming the sample into a standard Gaussian distribution, which is mathematically defined :

$$q(Y_k|Y_{k-1}) = \mathcal{N}(Y_k; \sqrt{1-\beta_k}Y_{k-1}, \beta_k I), \quad (1)$$

where $\beta_k \in (0,1)$ denotes the rescaled variance schedule that controls the magnitude of noise added at each step. To reduce the computational cost during training, this process can be simplified using properties of Gaussian transitions:

$$q(Y_k|Y_0) = \mathcal{N}(Y_k; \sqrt{\bar{\alpha}_k}Y_0, (1-\bar{\alpha}_k)I), \quad (2)$$

$$Y_k = \sqrt{\bar{\alpha}_k}Y_0 + \sqrt{(1-\bar{\alpha}_k)}\epsilon, \ \epsilon \sim \mathcal{N}(0, I), \quad (3)$$

where $\alpha_k = 1-\beta_k$, $\bar{\alpha}_k = \prod_{i=1}^{k}\alpha_i$, and ϵ represents a noise vector sampled from a standard Gaussian distribution. As the diffusion step k increases, Y_k progressively approximates the standard normal distribution $\mathcal{N}(0, I)$.

The reverse process aims to remove noise from Y_k, gradually recovering the original distribution Y_0 by learning reverse $p_\theta(Y_{k-1}|Y_k)$, which can be formulated as:

$$p_\theta(Y_{k-1}|Y_k) = \mathcal{N}(Y_{k-1}; \mu_\theta(Y_k, k, f), \Sigma_\theta(Y_k, k, f)), \quad (4)$$

where μ_θ and Σ_θ are neural networks predicting the mean and variance, respectively, and f is the conditional guidance.

Problem Definition

The trajectory prediction task entails estimating pedestrians' future positions within a scene based on their observed past movements. The model takes as input the 2D spatial coordinates of pedestrians over the observed time steps, denoted by $X_{1:t_{obs}} = \{X_t^i \in \mathbb{R}^2 | 1 \le t \le t_{obs}\}$, where $X_t^i = (x_t^i, y_t^i)$ represents the position of the ith pedestrian at time t. Similarly, the ground truth trajectory over the future time period is denoted as $Y_{1:t_{pred}} = \{Y_t^i \in \mathbb{R}^2 | 1 \le t \le t_{pred}\}$, and $Y_t^i = (x_t^i, y_t^i)$ represents the ground truth position of the ith pedestrian at time t in the future. The objective of this work is to predict the future trajectory $\hat{Y}^i$ of the ith pedestrian and to estimate its future positions $\hat{Y}_t^i = (\hat{x}_t^i, \hat{y}_t^i)$ at each time t.

Overview

The proposed architecture, illustrated in Figure 2, is a diffusion-based framework that models both long-term E and short-term I pedestrian motion intentions through dedicated modules with observed motion features F_{obs} as conditional guidance. During the trajectory generation process, a softmask classifier-free guidance mechanism is employed to adaptively integrate conditional signals and a refinement module estimates the residual error in the predicted noise.

Motion Encoder

The motion encoder extracts motion features from observed pedestrian trajectories while capturing social interactions. These features are used both for predicting intentions and conditioning the diffusion model for trajectory generation.

$$F_{obs} = \text{MotionEncoder}(X_{1:t_{obs}}) \in \mathbb{R}^{t_{obs}\times d}, \quad (5)$$

where $X \in \mathbb{R}^{t_{obs}\times 2}$ denotes the observed trajectories. We adopt the encoder from (Gu et al. 2022), proven effective in capturing complex pedestrian motion patterns.

Residual Polar Modeling for Intention Estimation

We propose a residual polar coordinate-based representation to model short-term pedestrian motion intention, inferred

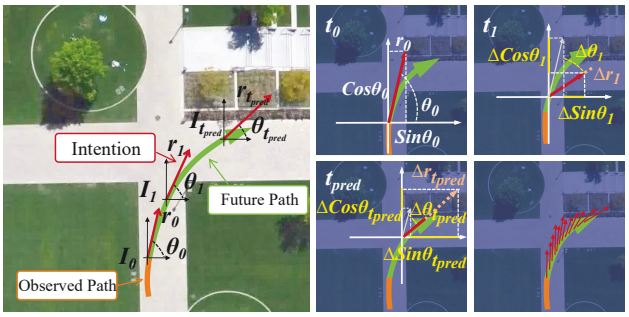

Figure 3: Illustration of short-term intention prediction. The sequence of short-term intentions is constructed using a polar-coordinate representation, as shown on the left. On the right, residual updates are recursively accumulated over time to refine these intention predictions.

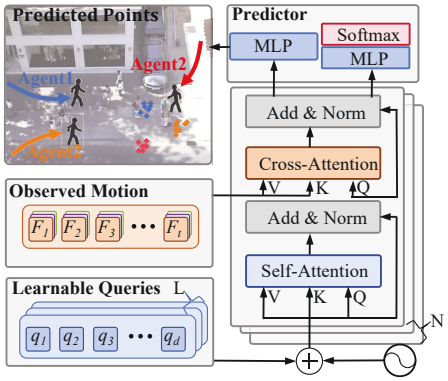

Figure 4: Illustration of the Endpoint Estimator. L learnable query tokens interact with observation features through cross-attention to generate multimodal endpoint predictions.

from observed motion features as shown in Figure 3. Instead of discretizing pedestrian intent into predefined motion classes (e.g., turning left or accelerating), intention at each future timestep t is represented in a continuous polar form as $I^i = \{\theta_t^i, r_t^i\}_{t=1}^{T_{\text{pred}}} \in \mathbb{R}^{T_{\text{pred}} \times 2}$, where $\theta_t^i \in (-\pi, \pi]$ denotes the directional angle, and $r_t^i \in \mathbb{R}^+$ denotes the magnitude of motion tendency. The origin point $\mathcal{O}_t^i$ corresponds to the pedestrian's current Cartesian position $\{x_i^t, y_i^t\}$. These components are computed as follows:

$$\theta_t^i = \arctan 2(a_t^{y_i}, a_t^{x_i}), r_t^i = \sqrt{(a_t^{x_i})^2 + (a_t^{y_i})^2 + \varepsilon}, \quad (6)$$

where $\arctan 2$ is the two-argument inverse tangent function, and $a_t^{x_i}$, $a_t^{y_i}$ denote the second-order temporal derivatives of the ith pedestrian's position in the lateral and longitudinal directions. A small constant ε is added to ensure numerical stability and avoid division by zero.

To capture the dynamic changes in future motion intentions, we adopt a transformer network to predict a sequence I^i of polar-coordinate representations. Instead of directly regressing the absolute values of the heading angle θ_t^i and magnitude r_t^i at each time step, we reformulate the prediction task into estimating the residual changes with respect to the previous state. This is motivated by predicting residuals simplifies the learning objective by focusing on local variations, which are typically smoother and less noisy than absolute values. The process is defined as:

$$\left[\Delta \cos \theta^i, \Delta \sin \theta^i, \Delta r^i\right] = \text{IntentPredictor}(F_{obs}). \quad (7)$$

We then convert the predicted residuals into scalar angular and magnitude increments, respectively: $\Delta\theta_t^i = \arctan 2(\Delta \sin \theta_t^i, \Delta \cos \theta_t^i)$ and $\Delta r_t^i = \text{softplus}(\Delta r_t^i) = \log(1 + \exp(\Delta r_t^i))$. The residual polar updates for the future intention I^i at each time step t are then computed by recursively accumulating these residuals over time:

$$\theta_t^i = \theta_0^i + \sum_{\tau=1}^{t} \Delta\theta_\tau^i, \quad r_t^i = r_0^i + \sum_{\tau=1}^{t} \Delta r_\tau^i, \quad (8)$$

where θ_0^i and r_0^i are initialised values based on the final frame of the observed trajectory $X_{t_{\text{obs}}}^i = \{x_{t_{obs}}^i, y_{t_{obs}}^i\}$.

This formulation incrementally refines intention predictions, ensuring smooth, consistent direction and magnitude estimates that better capture local motion dynamics.

Endpoints Prediction

The objective of destination prediction is to estimate the final endpoint of a pedestrian's trajectory, thereby capturing long-term behavioral intent based on observed motion history. To address the inherent multimodality of human behavior, our model predicts multiple candidate destinations. During training, the candidate closest to the ground truth is selected as supervision, while during inference, the destination with the highest confidence score is chosen.

In this task, we use a learnable query-based method with a transformer block to generate diverse destinations as shown in Figure 4. Specifically, we initialize L learnable endpoint query tokens $Q \in \mathbb{R}^{L \times d_q}$. These queries act as anchors representing plausible motion modes. To condition the queries on observed behavior, we refine them via the transformer's cross-attention using motion features $F_{obs} \in \mathbb{R}^{T_{\text{obs}} \times d}$.

For endpoint predictor, an MLP is applied to the refined multimodal motion features obtained from the transformer to regress a set of candidate endpoints, denoted as $E^i = \{e_1^i, e_2^i, \ldots, e_L^i\} \in \mathbb{R}^{L \times 2}$. Additionally, a separate MLP followed by a softmax layer is used to predict a confidence score for each candidate, forming a discrete probability distribution over the endpoint hypotheses: $P^i = \{p_1^i, p_2^i, \ldots, p_L^i\} \in \mathbb{R}^{L \times 1}$, where $\sum_{l=1}^{L} p_l^i = 1$. The endpoint prediction module is given as:

$$[E, P] = \text{EndpointPredictor}(F_{obs}, Q), \quad (9)$$

where EndpointPredictor is conditioned on observed motion features F_{obs} and learnable goal queries Q.

Condition-Guided Refinement

To guide the generation process toward intended motion semantics, we first adopt a modified classifier-free guidance strategy with dynamic masking, followed by a residual refinement step to correct prediction errors.

Instead of statically injecting conditions, we propose a dynamic soft-mask mechanism to modulate the contribution of each guidance signal. Specifically, a multi-layer perception layer followed by a sigmoid activation is used to learn

a soft weight M_m for each guidance source G_m, where $m \in \{o, s, l\}$ denotes the modality source: observed trajectory feature o, short-term intention s, and long-term goal l.

$$M_m = \text{Sigmoid}(\text{MLP}_m(G_m)), \tag{10}$$

$$G'_m = (1 - M_m) \cdot G_m + M_m \cdot \psi_m, \tag{11}$$

where ψ_m denotes a learnable token for masked guidance.

These soft-masked features G'_m are concatenated with the noised trajectory state Y_k and the corresponding diffusion step embedding E_k, then used to predict noise $\epsilon_k \in \mathbb{R}^{t_{obs} \times 2}$.

$$\epsilon_k = \text{NoiseEstimator}(\text{Concat}(G'_o, G'_s, G'_l, Y_k, E_k)). \tag{12}$$

Although the denoising network predicts the added noise ϵ_k, directly learning the full noise can be suboptimal due to its inherent stochasticity. To address this, we introduce a residual refinement module that explicitly learns to correct the residual error $\Delta\epsilon_k$ not captured by the main diffusion model. The final refined noise is computed as:

$$\epsilon_{\text{refined},k} = \epsilon_k + \Delta\epsilon_k, \tag{13}$$

where $\Delta\epsilon_k = \text{RefineNet}(G'_o, G'_s, G'_l, \epsilon_k, E_k) \in \mathbb{R}^{t_{obs} \times 2}$ denotes the predicted residual by the refinement network.

Training Optimization

To train the proposed method, we incorporate the short-term loss and the long-term loss in addition to the diffusion loss.

The short-term loss $\mathcal{L}_s$ supervises the intention estimator to more accurately approximate pedestrian motion tendencies within the observation context. Given the polar representation (θ, r) of intention, the angular component is trained using cosine distance to address its circular nature, while the magnitude is optimized via mean squared error.

$$\mathcal{L}_\theta = \frac{1}{T \times N} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(1 - \cos(\hat{\theta}_t^i - \theta_t^i)\right), \tag{14}$$

$$\mathcal{L}_r = \frac{1}{T \times N} \sum_{i=1}^{N} \sum_{t=1}^{T} \text{MSE}\left(\hat{r}_t^i - r_t^i\right), \tag{15}$$

$$\mathcal{L}_s = \lambda_\theta \cdot \mathcal{L}_\theta + \lambda_r \cdot \mathcal{L}_r, \tag{16}$$

where $\hat{\theta}_t^i$ and $\hat{r}_t^i$ are the predicted values, and θ_t^i and r_t^i are the corresponding ground truths. The loss weights λ_θ and λ_r are empirically set to 0.5 and 0.25.

The long-term loss guides the endpoint estimator to capture global motion patterns and accurately predict pedestrian destinations. Rather than optimizing over all candidate endpoints, which could dilute the learning signal due to the inherently multimodal nature of human motion, the endpoint loss $\mathcal{L}_e$ focuses on only minimizing the distance between the ground-truth endpoint and the closest predicted candidate.

$$\mathcal{L}_e = \frac{1}{N} \sum_{i=1}^{N} \min_{j \in L} \text{MSE}(\hat{e}_j^i, e^i), \tag{17}$$

where $\hat{e}^i$ and e^i denote the predicted and ground-truth endpoints, respectively.

Additionally, a negative log-likelihood loss $\mathcal{L}_p$ boosts the confidence of the closest goal, encouraging high probability for accurate predictions. To suppress incorrect candidates and enforce a sharper distribution, a penalty term is added.

$$\mathcal{L}_p = \frac{1}{N} \sum_{i=1}^{N} (-\log \hat{p}_{l^*}^i + \sum_{l \neq l^*} \log \hat{p}_l^i), \tag{18}$$

$$\mathcal{L}_l = \lambda_e \cdot \mathcal{L}_e + \lambda_p \cdot \mathcal{L}_p, \tag{19}$$

where $\hat{p}^i$ is the predicted probability over endpoint candidates, and l^* indicates the index of the predicted endpoint closest to the ground truth. The loss weights λ_e and λ_p are set to 1.0 and 0.5, respectively.

The diffusion loss effectively guides the noise estimator to generate reliable predictions Y_k for use during inference, and is subsequently integrated with both the short intention and goal objectives to form the final weighted training loss.

$$\mathcal{L}_{dif} = \mathbb{E}_k \left\| \epsilon_k - \hat{\epsilon}_{refined,k} \right\|^2, \tag{20}$$

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_l + \lambda_{dif} \cdot \mathcal{L}_{dif}, \tag{21}$$

where the diffusion weight λ_{dif} is empirically set to 1.

Inference

The reverse diffusion process is typically modeled as a conditional Markov chain, which gradually denoises inputs to generate trajectories. However, its stochasticity and iterative nature incur high computational costs and may cause deviations from ground truth. To mitigate this, we adopt the deterministic DDIM sampling strategy, which removes randomness and reduces inference steps. Under this method, a sample $Y_{k-\gamma}$ is deterministically derived from Y_k as:

$$\begin{aligned} Y_{k-\gamma} = \sqrt{\alpha_{k-\gamma}} \left(\frac{Y_k - \sqrt{1-\alpha_k}\, \epsilon_{refined}}{\sqrt{\alpha_k}} \right) \\ + \sqrt{1-\alpha_{k-\gamma}}\, \epsilon_{refined}. \end{aligned} \tag{22}$$

Experiments

Experimental Settings

Datasets: We evaluate the model on four pedestrian trajectory datasets: ETH (Pellegrini et al. 2009), UCY (Alon Lerner 2007), Stanford Drone Dataset (SDD) (Robicquet et al. 2016), and NBA Dataset. SDD is a large-scale dataset captured from a bird's-eye view using drone cameras. The ETH dataset comprises ETH and HOTEL, while the UCY dataset includes ZARA1, ZARA2, and UNIV. The NBA SportVU dataset tracks the trajectories of players and the ball during NBA basketball games. All sequences consist of 8 observed and 12 predicted frames over 8 seconds.

Metrics: Following previous works, we adopt two standard evaluation metrics. Average Displacement Error (ADE) measures the mean Euclidean distance between the predicted and ground-truth trajectories over all prediction time steps, while Final Displacement Error (FDE) computes the Euclidean distance between the predicted final position and the ground-truth endpoint. In line with prior studies, we report the best result among 20 sampled trajectories.

Method	Venue	Year	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Social-GAN	CVPR	2018	0.87 / 1.62	0.67 / 1.37	0.76 / 1.52	0.35 / 0.68	0.42 / 0.84	0.61 / 1.21
Social-STGCNN	CVPR	2020	0.64 / 1.11	0.49 / 0.85	0.44 / 0.79	0.34 / 0.53	0.30 / 0.48	0.44 / 0.75
Social-VAE	ECCV	2022	0.41 / 0.58	0.13 / 0.19	0.21 / 0.36	0.17 / 0.29	0.13 / 0.22	0.21 / 0.33
MID	CVPR	2022	0.39 / 0.66	0.13 / 0.22	0.22 / 0.45	0.17 / 0.30	0.13 / 0.27	0.21 / 0.38
LED	CVPR	2023	0.39 / 0.58	**0.11** / 0.17	0.26 / 0.43	0.18 / 0.26	0.13 / 0.22	0.21 / 0.33
Graph-TERN	AAAI	2023	0.42 / 0.58	0.14 / 0.23	0.26 / 0.45	0.21 / 0.37	0.17 / 0.29	0.24 / 0.88
EigenTrajectory	ICCV	2023	0.36 / 0.53	0.12 / 0.19	0.24 / 0.43	0.19 / 0.33	0.14 / 0.24	0.21 / 0.34
TUTR	ICCV	2023	0.40 / 0.61	**0.11** / 0.18	0.23 / 0.42	0.18 / 0.34	0.13 / 0.25	0.21 / 0.36
SMEMO	TPAMI	2024	0.39 / 0.59	0.14 / 0.20	0.23 / 0.41	0.19 / 0.32	0.15 / 0.26	0.22 / 0.35
HighGraph	CVPR	2024	0.40 / 0.55	0.13 / 0.17	**0.20** / **0.33**	0.17 / 0.27	**0.11** / 0.21	0.20 / **0.30**
PPT	ECCV	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
MoFlow	CVPR	2025	0.40 / 0.57	**0.11** / 0.17	0.23 / 0.39	**0.15** / 0.26	0.12 / 0.22	0.20 / 0.32
Ours	–	–	**0.34** / 0.52	0.15 / 0.24	**0.20** / 0.36	**0.15** / **0.24**	**0.11** / **0.20**	**0.19** / 0.31

Table 1: Quantitative comparisons with state-of-the-art methods on the ETH/UCY dataset. Bold numbers indicate the best performance. Underlined numbers denote the second-best results.

Time	Social-GAN CVPR/2018	Social-STGCNN CVPR/2020	PECNet ECCV/2020	NPSN CVPR/2022	GroupNet CVPR/2022	MID CVPR/2022	LED CVPR/2023	MoFlow CVPR/2025	Ours
1.0s	0.41/0.62	0.34/0.48	0.40/0.71	0.35/0.58	0.26/0.34	0.28/0.37	**0.18**/0.27	**0.18/0.25**	0.21/**0.25**
2.0s	0.81/1.32	0.71/0.94	0.83/1.61	0.68/1.23	0.49/0.70	0.51/0.72	0.37/0.56	0.34/0.47	**0.32/0.45**
3.0s	1.19/1.94	1.09/1.77	1.27/2.44	1.01/1.76	0.73/1.02	0.71/0.98	0.58/0.84	0.52/**0.67**	**0.48**/0.75
Total (4.0s)	1.59/2.41	1.53/2.26	1.69/2.95	1.31/1.79	0.96/1.30	0.96/1.27	0.81/1.10	**0.71/0.87**	0.79/0.99

Table 2: Quantitative comparisons on NBA dataset. Bold indicates the best results, while underlined indicates the second-best.

Implementation: The network is implemented in PyTorch. The noise estimation module uses 4 Transformer layers with hidden size 512 and 4 attention heads. The short-term intention estimator employs 4 self-attention layers mapping features to 256 dimensions, while the long-term endpoint estimator adopts a similar structure with cross-attention to capture trajectory-level context. We set diffusion steps to $K = 100$ and apply DDIM sampling with stride 20. The model is trained using Adam with learning rate 0.001 and batch size 256. All experiments are conducted on NVIDIA RTX 5090 GPUs and Intel Xeon 8481C CPUs.

Quanitative Evaluation

Compared methods: MoFlow (Fu et al. 2025), PPT (Lin et al. 2024), HighGraph (Kim et al. 2024), SMEMO (Marchetti et al. 2024), TUTR (Shi et al. 2023), EigenTrajectory (Bae, Oh, and Jeon 2023), Graph-TERN (Bae and Jeon 2023), LED (Mao et al. 2023), MID (Gu et al. 2022), Social-VAE (Xu, Hayet, and Karamouzas 2022), Social-STGCNN (Mohamed et al. 2020), Sophie (Sadeghian et al. 2019), Social-GAN (Gupta et al. 2018), PECNet (Mangalam et al. 2020), NPSN (Bae, Park, and Jeon 2022), GroupNet (Xu et al. 2022), and PCCSNet (Sun et al. 2021).

The quantitative evaluation results on the UCY/ETH datasets are summarized in Table 1. Despite variations across individual subsets, our method consistently delivers competitive performance. Specifically, it achieves the lowest ADE on 4 out of 5 datasets and ranks either first or second in FDE on 4 of them. On average, the ADE is reduced from 0.20 to 0.19. In the ETH scenario, our method attains the best ADE, lowering it from 0.36 to 0.34. Similarly, for the Zara1 subset, the ADE is reduced from 0.17 to 0.15, and the FDE from 0.27 to 0.24. On the NBA dataset Table 2, our method attains the lowest ADE at 2 s and 3 s, as well as the lowest FDE at the 1 s and 2 s horizons. Additionally, as shown in Table 3, the proposed method achieves competitive results on the SDD dataset, with ADE reduced from 7.03 to 6.85 and the second-best FDE.

Methods	Venue	Year	ADE	FDE
Social-GAN	CVPR	2018	27.23	41.44
Sophie	CVPR	2019	16.27	29.38
PECNet	ECCV	2020	9.96	15.88
PCCSNet	ICCV	2021	8.62	16.16
Social-VAE	ECCV	2022	8.10	11.72
MID	CVPR	2022	7.61	14.30
LED	CVPR	2023	8.48	11.66
TUTR	ICCV	2023	7.76	12.69
PPT	ECCV	2024	7.03	**10.65**
MoFlow	CVPR	2025	7.50	11.96
Ours	–	–	**6.85**	11.22

Table 3: Comparisons with state-of-the-art methods on the SDD dataset. Text in bold numbers denotes the best result.

Qualitative Evaluation

To qualitatively evaluate our model, we visualize representative cases and compare with MID and Social-VAE.

As shown in Figure 5, which visualizes four examples per dataset, our method consistently produces trajectories that align more closely with the ground truth than competing approaches. For instance, in the first two ETH cases involving subtle turns, all methods capture the motion trend, but ours more accurately follows the true path. In the first ZARA ex-

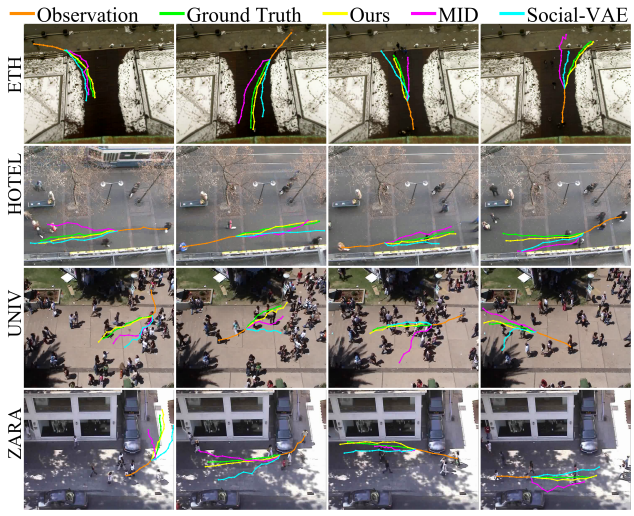

Figure 5: Visualisation of prediction results on the ETH/UCY dataset. Our method (yellow) is compared with MID (purple) and Social-VAE (cyan) across four scenarios from ETH and UCY datasets.

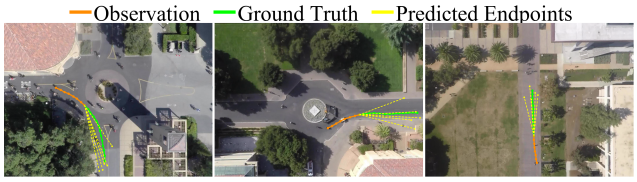

Figure 6: Illustration of predicted endpoints on three scenarios from the SDD dataset.

ample, where a sharp turn occurs, only our method adapts well, producing a precise trajectory. Additionally, Figure 6 illustrates predicted endpoint candidates on the SDD dataset. While some predictions deviate, our model reliably includes candidates near the true goal, offering strong guidance for accurate future trajectory generation.

Ablation Study

To assess the impact of different model components, we conduct ablation studies on the following aspects.

As shown in Table 4, removing either the short or long intention branch degrades performance, indicating that both levels of intention modeling are crucial for accurate prediction. In contrast, omitting the softmask or residual noise refinement yields smaller yet consistent drops, suggesting that these modules mainly stabilize the diffusion process and modulate intention conditioning. Overall, the results suggest that intention modeling drives most of the gains, with architectural refinements playing a complementary role.

Similarly, an ablation study on the number of candidate endpoints M was conducted, as shown in Table 5. Results indicate that prediction performance is sensitive to M, and the relationship is not linear. Overall performance across scenarios is maximized when M is 5. A possible explanation is that too many candidates may dilute the model's focus, making it harder to capture a coherent and interpretable intention distribution, while too few may limit its ability to represent the diversity of plausible intent hypotheses.

We further evaluate the impact of diffusion steps K by varying it from 10 to 200, with detailed results shown in Table 6. Overall prediction performance consistently peaks at $K = 100$, where both ADE and FDE reach their lowest values. Too few steps may hinder the model's ability to effectively capture fine-grained noise transitions, while too many make subtle differences between steps harder to distinguish, potentially exceeding the model's representational capacity.

Long	Short	Softmask	Refine	ETH & UCY	SDD
✓	✓	✓	✓	**0.19 / 0.31**	**6.85/ 11.22**
✓	✓	✓		0.23 / 0.35	7.92/ 12.45
✓	✓		✓	0.25 / 0.41	9.25/ 13.82
✓		✓	✓	0.31 / 0.38	8.96/ 13.12
	✓	✓	✓	0.29 / 0.36	8.91/ 12.79

Table 4: Ablation study on model components on the ETH, UCY, and SDD. Bold is the best.

M	1	3	5	10	20
ETH	0.39/0.61	0.36/0.56	0.34/**0.52**	**0.33/0.52**	0.40/0.59
HOTEL	0.18/0.31	**0.15**/0.28	**0.15/0.24**	0.18/0.25	0.21/0.26
UNIV	0.25/0.39	0.21/**0.35**	**0.20**/0.36	**0.20**/0.41	0.23/0.42
ZARA1	0.20/0.29	0.18/0.27	**0.15/0.24**	0.16/0.30	0.24/0.32
ZARA2	0.16/0.29	**0.11**/0.24	**0.11/0.20**	0.17/**0.20**	0.29/0.32
AVG	0.24/0.38	0.20/0.34	**0.19/0.31**	0.21/0.34	0.27/0.38

Table 5: Ablation study on the number of endpoint candidates M on the ETH and UCY datasets. Bold is the best.

K	10	50	100	150	200
ETH	0.37/0.58	**0.34**/0.53	**0.34/0.52**	0.36/0.60	0.39/0.65
HOTEL	0.16/0.29	0.16/**0.23**	**0.15**/0.24	0.18/0.25	0.21/0.27
UNIV	0.23/0.38	**0.20**/0.38	**0.20/0.36**	0.23/**0.36**	0.23/0.42
ZARA1	0.19/0.29	0.17/0.26	0.15/**0.24**	**0.14**/0.25	0.16/0.25
ZARA2	0.13/0.22	**0.11**/0.22	**0.11/0.20**	0.12/0.23	0.14/0.23
AVG	0.22/0.35	0.20/0.32	**0.19/0.31**	0.21/0.34	0.23/0.36

Table 6: Ablation study on diffusion step K on the ETH and UCY datasets. Bold is the best.

Conclusion

In this work, we propose a diffusion-based trajectory prediction framework enhanced by long- and short-term intentions. Short-term motion is represented in polar coordinates for precise local modeling, while long-term intention is captured via predicted endpoints providing global cues. Softmask-based classifier-free guidance and residual noise estimation improve generation quality by enhancing error alignment during denoising. Experiments demonstrate competitive performance against state-of-the-art methods.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant No. 2024YFB4710900, the National Natural Science Foundation of China (NSFC) under Grant No. U24A20265, the Science, Technology, and Innovation Commission of Shenzhen Municipality, China, under Grant No. ZDSYS20220330161800001, JCYJ20240813094212017, the Shenzhen Science and Technology Program under Grant No. KQTD20221101093557010, the Guangdong Science and Technology Program under Grant No. 2024B1212010002.

References

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 961–971.

Alon Lerner, D. L., Yiorgos Chrysanthou. 2007. Crowds by example. *Computer Graphics Forum*, 26(3): 655–664.

Bae, I.; and Jeon, H.-G. 2023. A Set of Control Points Conditioned Pedestrian Trajectory Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(5): 6155–6165.

Bae, I.; Oh, J.; and Jeon, H.-G. 2023. EigenTrajectory: Low-Rank Descriptors for Multi-Modal Trajectory Forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9983–9995.

Bae, I.; Park, J.-H.; and Jeon, H.-G. 2022. Non-Probability Sampling Network for Stochastic Human Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6467–6477.

Bae, I.; Park, Y.-J.; and Jeon, H.-G. 2024. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17890–17901.

Chiara, L. F.; Coscia, P.; Das, S.; Calderara, S.; Cucchiara, R.; and Ballan, L. 2022. Goal-driven self-attentive recurrent networks for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2518–2527.

Choi, Y.; Mercurius, R. C.; Mohamad Alizadeh Shabestary, S.; and Rasouli, A. 2024. DICE: Diverse Diffusion Model with Scoring for Trajectory Prediction. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, 3023–3029.

Duan, J.; Wang, L.; Long, C.; Zhou, S.; Zheng, F.; Shi, L.; and Hua, G. 2022. Complementary Attention Gated Network for Pedestrian Trajectory Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1): 542–550.

Eiffert, S.; Kong, H.; Pirmarzdashti, N.; and Sukkarieh, S. 2020. Path Planning in Dynamic Environments using Generative RNNs and Monte Carlo Tree Search. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 10263–10269.

Firl, J.; Stübing, H.; Huss, S. A.; and Stiller, C. 2012. Predictive maneuver evaluation for enhancement of Car-to-X mobility data. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, 558–564.

Fu, Y.; Yan, Q.; Wang, L.; Li, K.; and Liao, R. 2025. Moflow: One-step flow matching for human trajectory forecasting via implicit maximum likelihood estimation based distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17282–17293.

Gu, J.; Sun, C.; and Zhao, H. 2021. DenseTNT: End-to-end Trajectory Prediction from Dense Goal Sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15283–15292.

Gu, T.; Chen, G.; Li, J.; Lin, C.; Rao, Y.; Zhou, J.; and Lu, J. 2022. Stochastic Trajectory Prediction via Motion Indeterminacy Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17092–17101.

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/CVF Conference on Computer Vision and Pattern Recognition*, 2255–2264.

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

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Kim, S.; Chi, H.-g.; Lim, H.; Ramani, K.; Kim, J.; and Kim, S. 2024. Higher-order Relational Reasoning for Pedestrian Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15251–15260.

Klingelschmitt, S.; Platho, M.; Groß, H.-M.; Willert, V.; and Eggert, J. 2014. Combining behavior and situation information for reliably estimating multiple intentions. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, 388–393.

Kothari, P.; Sifringer, B.; and Alahi, A. 2021. Interpretable Social Anchors for Human Trajectory Forecasting in Crowds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15551–15561.

Li, K.; Eiffert, S.; Shan, M.; Gomez-Donoso, F.; Worrall, S.; and Nebot, E. 2021. Attentional-GCNN: Adaptive Pedestrian Trajectory Prediction towards Generic Autonomous Vehicle Use Cases. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 14241–14247.

Li, X.; Thickstun, J.; Gulrajani, I.; Liang, P. S.; and Hashimoto, T. B. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35: 4328–4343.

Liao, H.; Wang, C.; Zhu, K.; Ren, Y.; Gao, B.; Li, S. E.; Xu, C.; and Li, Z. 2025. Minds on the move: Decoding trajectory prediction in autonomous driving with cognitive

insights. *IEEE Transactions on Intelligent Transportation Systems*.

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

Liu, Y.; Zhang, Y.; Li, K.; Qiao, Y.; Worrall, S.; Li, Y.-F.; and Kong, H. 2023. Knowledge-aware Graph Transformer for Pedestrian Trajectory Prediction. In *Proceedings of the International Conference on Intelligent Transportation Systems*, 4360–4366.

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 *Proceedings of the European Conference on Computer Vision*, 759–776.

Mao, W.; Xu, C.; Zhu, Q.; Chen, S.; and Wang, Y. 2023. Leapfrog Diffusion Model for Stochastic Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5517–5526.

Marchetti, F.; Becattini, F.; Seidenari, L.; and Bimbo, A. D. 2024. SMEMO: Social Memory for Trajectory Forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6): 4410–4425.

Mohamed, A.; Qian, K.; Elhoseiny, M.; and Claudel, C. 2020. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14424–14432.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning*, 8162–8171.

Pellegrini, S.; Ess, A.; Schindler, K.; and van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 261–268.

Rempe, D.; Luo, Z.; Peng, X. B.; Yuan, Y.; Kitani, K.; Kreis, K.; Fidler, S.; and Litany, O. 2023. Trace and Pace: Controllable Pedestrian Animation via Guided Trajectory Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13756–13766.

Robicquet, A.; Sadeghian, A.; Alahi, A.; and Savarese, S. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Proceedings of the European Conference on Computer Vision*, 549–565.

Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; and Savarese, S. 2019. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1349–1358.

Shi, L.; Wang, L.; Zhou, S.; and Hua, G. 2023. Trajectory Unified Transformer for Pedestrian Trajectory Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9641–9650.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the International Conference on Machine Learning*, 2256–2265.

Sun, J.; Li, Y.; Fang, H.-S.; and Lu, C. 2021. Three Steps to Multimodal Trajectory Prediction: Modality Clustering, Classification and Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13230–13239.

Wang, X.; Su, T.; Da, F.; and Yang, X. 2023. Prophnet: Efficient agent-centric motion forecasting with anchor-informed proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21995–22003.

Wang, Z.; Miao, H.; Wang, S.; Wang, R.; Wang, J.; and Zhang, J. 2025. C2f-tp: A coarse-to-fine denoising framework for uncertainty-aware trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 12810–12817.

Wu, K.; Zhou, Y.; Shi, H.; Li, X.; and Ran, B. 2024. Graph-Based Interaction-Aware Multimodal 2D Vehicle Trajectory Prediction Using Diffusion Graph Convolutional Networks. *IEEE Transactions on Intelligent Vehicles*, 9(2): 3630–3643.

Xing, Z.; Zhang, X.; Hu, Y.; Jiang, B.; He, T.; Zhang, Q.; Long, X.; and Yin, W. 2025. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1602–1611.

Xu, C.; Li, M.; Ni, Z.; Zhang, Y.; and Chen, S. 2022. GroupNet: Multiscale Hypergraph Neural Networks for Trajectory Prediction with Relational Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6488–6497.

Xu, P.; Hayet, J.-B.; and Karamouzas, I. 2022. SocialVAE: Human Trajectory Prediction using Timewise Latents. In *Proceedings of the European Conference on Computer Vision*, 511–528.

Yang, B.; Wei, Z.; Hu, C.; Cai, Y.; Wang, H.; and Hu, H. 2024a. Real-Time Pedestrian Crossing Anticipation Based on an Action–Interaction Dual-Branch Network. *IEEE Transactions on Intelligent Transportation Systems*, 25(12): 21021–21034.

Yang, B.; Zhu, J.; Hu, C.; Yu, Z.; Hu, H.; and Ni, R. 2024b. Faster Pedestrian Crossing Intention Prediction Based on Efficient Fusion of Diverse Intention Influencing Factors. *IEEE Transactions on Transportation Electrification*, 10(4): 9071–9087.

Yin, Z.; Lai, T.; Barcelos, L.; Jacob, J.; Li, Y.; and Ramos, F. 2025. Diverse Motion Planning with Stein Diffusion Trajectory Inference. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 15610–15616.

Zhao, H.; Gao, J.; Lan, T.; et al. 2021. Tnt: Target-driven trajectory prediction. In *Proceedings of the Conference on Robot Learning*, 895–904.