DifTraj: Diffusion Inspired by Intrinsic Intention and Extrinsic Interaction for Multi-Modal Trajectory Prediction

Yanghong Liu , **Xingping Dong*** , **Yutian Lin** and **Mang Ye**
School of Computer Science, Wuhan University
liuyanghong0416@gmail.com, {yutian.lin, yemang, xingpingdong}@whu.edu.cn

Abstract

Recent years have witnessed the success of generative adversarial networks and diffusion models in multi-model trajectory prediction. However, prevailing algorithms only explicitly consider human interactions but ignore the modeling of human intentions, yielding that the generated results deviate largely from real trajectories in complex scenes. In this paper, we analyze the conditions of multi-modal trajectory prediction from two objective perspectives and propose a novel end-to-end framework based on the diffusion model to predict more precise and socially-acceptable trajectories. Firstly, a spatial-temporal aggregation module is built to extract the extrinsic interaction features for capturing socially-acceptable behaviors. Secondly, we explicitly construct the intrinsic intention module to obtain intention features for precise prediction. Finally, we estimate a noise trajectory distribution with these two features as the initiation of the diffusion model and leverage the denoising process to obtain the final trajectories. Furthermore, to reduce the noise of the initiative trajectory estimation, we present a novel *sample consistency loss* to constrain multiple predictions. Extensive experiments demonstrate that our method outperforms the state-of-the-art methods on ETH-UCY and SDD benchmarks, specifically achieving 19.0%/24.2% ADE/FDE improvement on ETH-UCY.

1 Introduction

Human trajectory prediction serves a significant role in vast applications, such as intelligent transportation system [Qiao *et al.*, 2014; Teeti *et al.*, 2022], human-robot system [Cheng *et al.*, 2020], and surveillance system [Valera and Velastin, 2005]. It is a challenge to build an effective model for trajectory prediction since trajectories are inherently stochastic. A reasonable solution to the uncertainty in stochastic trajectories is to generate multiple trajectory distributions. Some previous works [Su *et al.*, 2017; Gupta *et al.*, 2018; Mangalam *et al.*, 2020; Bae *et al.*, 2022] have emerged great potentials from the generative models, such as generative adversarial network (GAN) [Goodfellow *et al.*, 2014] and conditional variational auto-encoder (CVAE) [Sohn *et al.*, 2015]. However, GAN often suffers from unstable training, and the prediction results of CVAE have significant biases from the actual trajectories. Recently, inspired by the success of diffusion model [Ho *et al.*, 2020] in the field of image [Nichol and Dhariwal, 2021] and video [Kong *et al.*, 2020], MID [Gu *et al.*, 2022] utilizes the standard diffusion model to produce multiple trajectory distribution. Compared with other generative models, it obtains a more stable training process and reliable trajectories. To accelerate the inference process, LED [Mao *et al.*, 2023] directly learns sophisticated trajectory distributions as the leapfrog initializer of the diffusion model. However, most existing methods only consider humans' extrinsic interaction, resulting in a limited set of typical trajectories that can only represent part of the range of possible outcomes.

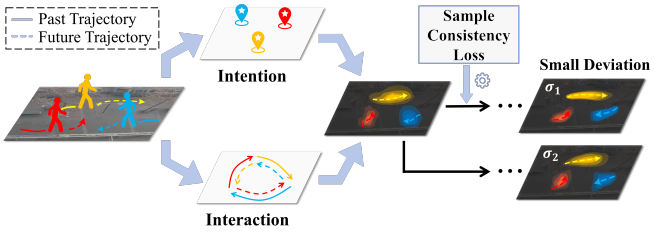

Figure 1: Our DifTraj achieves alleviating the indeterminacy by learning incorporated with interaction and intention features, and generates multiple trajectories based on the diffusion model with the end-to-end training. To reduce the noise of the initiative trajectory estimation, we present a novel *sample consistency loss* to constrain multiple predictions.

To achieve high-precision multi-modal trajectory prediction, we rethink the decision-making process of pedestrians and divide the factors into two aspects: internal and external. Firstly, pedestrians usually keep a walking goal in the mind, which acts on the entire body to control the direction and speed of movements. Essentially, it can be regarded as a purposeful behavior emitted by the pedestrian internally [Xu *et al.*, 2018]. Moreover, external interactions between different entities in the environment, such as pedestrians avoiding collision or going together, can lead to unpredictable and

*Corresponding author.

complex trajectories [Shu *et al.*, 2018]. For example, if a pedestrian falls abruptly, the following may also stop and offer assistance, which affects the followers' future movements. Jointly considering the impacts of intrinsic intention and extrinsic interaction to build a trajectory prediction model is still challenging, especially for modeling intention. Most existing methods do not explicitly model the intention but implicitly predict the entire future trajectory containing the intention estimation.

In this paper, we aim to propose a novel framework of multi-modal trajectory prediction to simulate the decision-making process of humans by utilizing intrinsic intentions and extrinsic interactions shown in Figure 1. Our framework contains three key components: interaction feature extracting, intention modeling, and trajectory inference. 1) **Interaction feature extracting**: We classify the influence of interactions into two aspects: temporal and spatial-wise. In detail, we propose a time-attention module to grasp the temporal dependence of each past timestamp and leverage graph attention to learn spatial interactions among different agents. 2) **Intention modeling**: Due to the difficulty in detecting intention signals inside the mind, we use the destination point to represent the latent intention. Specifically, we present a novel module based on attention for extracting the intention features from the observed trajectory. 3) **Trajectory inference**: Two modality features are fused to form an initially noised trajectory distribution. Then, a diffusion model will complete the denoising process from coarse to fine. Notably, our framework can generate multiple socially-acceptable trajectories close to realistic distribution by fully considering the intrinsic intention and extrinsic interaction. Furthermore, to improve the accuracy of noised-trajectory estimation and reduce denoising steps, we propose a new *sample consistency loss* to constrain multiple predictions by forcing all prediction samples to approach the ground truth. Sufficient experiments have been conducted on two public datasets, ETH-UCY and SDD, demonstrating that our method outperforms the previous SOTA methods by a large margin. For example, our test performance on ETH-UCY is improved by 19.0% and 24.2% in terms of ADE and FDE metrics, respectively.

Our contributions are summarised as follows.

- We propose a novel end-to-end paradigm based on the diffusion model for multiple trajectory prediction. By learning from the intrinsic intention and extrinsic interaction, our framework can reduce indeterminacy and predict socially-acceptable trajectories for humans.
- A novel intention module based on attention is presented to capture intention signals inside the human mind. This is the first work to explicitly build an intention model by destination points in human trajectory prediction.
- We introduce a loss constraint named *sample consistency loss* for multi-modal trajectory prediction to implicitly supervise the model's ability to learn diverse samples' distribution with low variance, meanwhile improving prediction performance.
- Extensive experiments have demonstrated that our proposed method achieves state-of-the-art performance on several human trajectory benchmarks, i.e., ETH-UCY and SDD. Compared with the previous SOTA methods, the test performance is improved by 19.0%/24.2% regarding the ADE/FDE metrics on ETH-UCY.

2 Method

2.1 Problem Formulation

Trajectory prediction constitutes a time-series regression task that forecasts the future movements of one or multiple agents based on observed movements. A commonly employed technique for mathematically representing trajectories involves sampling pedestrian data in video frames at a fixed frequency and obtaining a series of 2D trajectory points through coordinate transformation. Considering the trajectory of an agent i changing continuously with time, (x_t^i, y_t^i) indicates the position of the agent i. The set of past trajectory is defined as $\mathbf{X} = \{X^i | X^i = (x_t^i, y_t^i), i \in (1, \ldots, N)\}$ from time steps $t = 1, \ldots, t_{obs}$ and the ground truth can be defined similarly as $\mathbf{Y} = \{Y^i | Y^i = (x_t^i, y_t^i), i \in (1, \ldots, N)\}$ from time steps $t = t_{obs} + 1, \ldots, t_{pred}$, N is total number of pedestrian. Denote $\Omega^i = (x_t^i, y_t^i)|_{t=t_{pred}}$ as a destination point of pedestrian i, which is also referred to the intention of the current trajectory sequence. The predicted future trajectories of all agents are presented as $\widehat{\mathbf{Y}} = [\widehat{Y}^1, \ldots, \widehat{Y}^N]$. Usually, different social behaviors may appear in different scenes, e.g., walking alone and merging into a group. In this work, we split the number of pedestrians appearing in different scenes to better model their interactions.

2.2 Framework Overview

As shown in Figure 2, DifTraj consists of three main parts: interaction feature extracting, intention modeling, and trajectory inference. Multi-modality features learned from humans' intrinsic intention and extrinsic interaction are utilized to guide the estimation of trajectory mean and variance, and then reparameterization techniques are applied to initialize the noised distribution. Finally, the denoising process from coarse to fine is conducted using the diffusion model.

2.3 Interaction Feature Extracting

Modeling human interaction is one essential technique in the trajectory prediction [Xu *et al.*, 2022a; Makansi *et al.*, 2022]. However, the interaction model in existing diffusion-based methods is inadequate in capturing temporal and spatial interaction features. For example, LED [Mao *et al.*, 2023] employs a transformer-based module to capture social influences. However, this module only focuses on positional information in the sequence and cannot directly handle spatial relationships between agents.

Due to the interaction along the temporal and spatial dimensions, a comprehensive interaction influence should attempt to incorporate the temporal and spatial interaction features. We utilize LSTM [Hochreiter and Schmidhuber, 1997] to explicitly model the temporal correlations between interactions, where all cells in LSTM can share hidden states from diverse pedestrians. To grasp the temporal dependence of each past timestamp, we design a time-attention module corresponding to the channel-attention in the SE layer [Hu *et al.*,

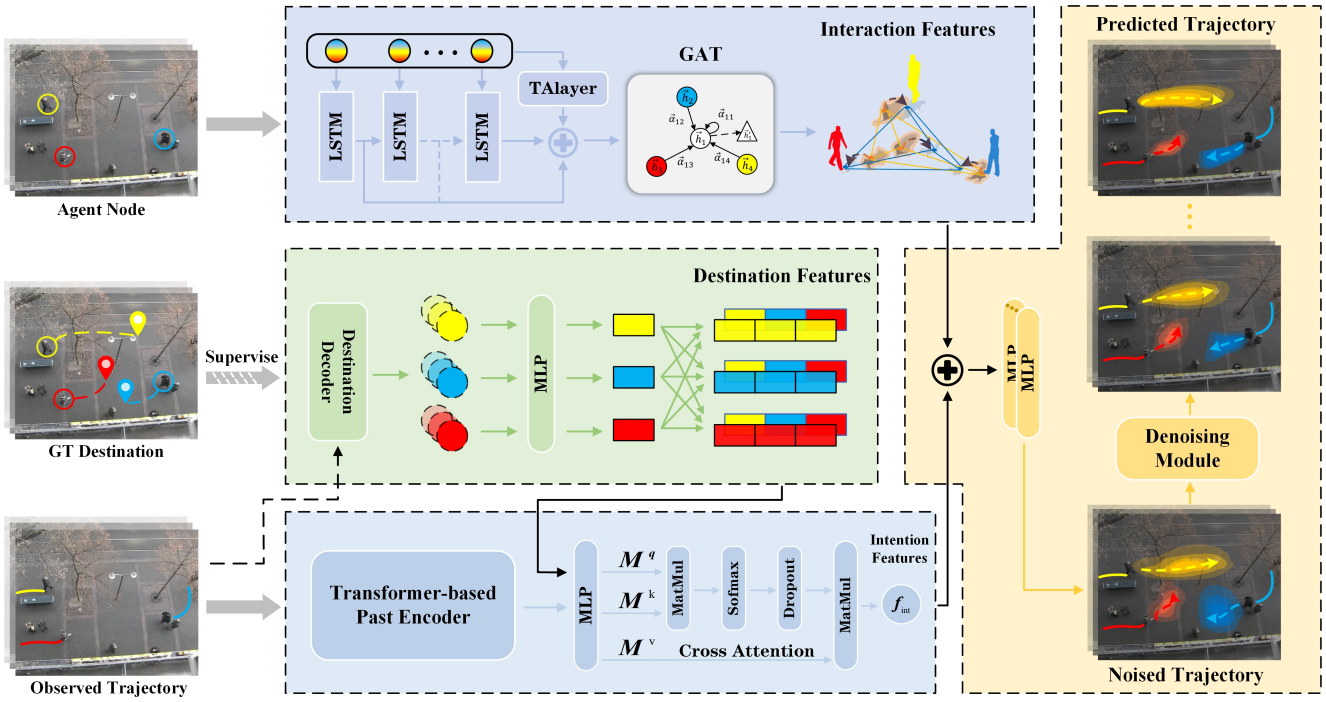

Figure 2: The framework of our method DifTraj, which incorporates multi-modality features learning from the intrinsic intention and extrinsic interaction to generate noised trajectories as the input of denoising process instead of traditionally adding Gaussian noise into original trajectory data, accelerating inference speed and improving the precision of multi-modal prediction. Note that MLP indicates the multilayer perceptron and GAT is the graph attention module.

2018]. This design ensures that the past trajectory of each timestamp is carefully encoded and measured. Subsequently, temporal interaction features are given different weights on time dimension. For more details, please refer to Figure **??** in the supplementary material.

$$e_t^i = \phi(x_t^i, y_t^i; W_{emb}), \tag{1}$$

$$h_t^i, c_t^i = \text{LSTM}(h_{t-1}^i, e_t^i; W_l), \tag{2}$$

$$f_{ext}^{tem} = \sum_{i=1}^{N} \sum_{t=1}^{T_{obs}} h_t^i, \tag{3}$$

$$\mathbf{z}_{tem} = \text{TAlayer}(\mathbf{x}, \mathbf{y}), \tag{4}$$

where ϕ is the embedding function, W_{emb} is a learnable weight matrix. Temporal interaction features f_{ext}^{tem} are composed of the hidden states h_t^i from the LSTM. The operator TA layer assigns different weights to the original position on the time dimension.

We feed the hidden states encoded by LSTM into the GAT [Veličković *et al.*, 2017] module to aggregate information from neighbors by assigning different importance to different nodes. The number of pedestrians in each scene is counted, and we split and pack the number of pedestrians that appeared in the same scenes like $seq = [start, end]$. Consequently, GAT operates on a dynamic number of pedestrians and outputs the aggregated spatial interaction states from other neighbors in the same scene. Eventually, we complete the fusion of spatial and temporal information to form extrinsic interaction feature f_{ext} by concatenating f_{ext}^{tem} and f_{ext}^{spa}:

$$f_{ext}^{spa} = \sum_{seq=1}^{M} \text{GAT}(\mathbf{e}^{seq}, [\mathbf{h}_{lstm}^{seq}; \mathbf{z}_{tem}^{seq}]), \tag{5}$$

$$f_{ext} = f_{ext}^{tem} \oplus f_{ext}^{spa}, \tag{6}$$

where spatial interaction features f_{ext}^{spa} are obtained by encoding $\mathbf{e}^{seq}$, $\mathbf{h}_{lstm}^{seq}$ and $\mathbf{z}_{tem}^{seq}$ with the GAT module.

2.4 Intention Modeling

In reality, humans are goal-driven [Tran *et al.*, 2021], and the intrinsic intention can be represented as the end point of trajectories (i.e., destination). Modeling destinations will continuously provide more accurate and specific messages for future trajectory prediction. However, although the dataset provides the future ground truth (GT) of trajectories, the intention of the destination in the real world is often unavailable and is a part of the prediction. Therefore, the model is required to combine past observations to infer the distribution of pedestrian destinations. During the training phase, we first model past observations and raw destinations. Then, we set an attention-based module to improve the ability to map past observations into future intentions. The model can generate coarse destinations by only inputting past trajectories during the inference phase by training a destination decoder. The ground truth of the destination is utilized to supervise the training process of the decoder in Figure 2. The estimated destinations are obtained as follows:

$$\widehat{\mathbf{\Omega}} = \text{Decoder}_{des}(\mathbf{X}). \tag{7}$$

However, this ambiguous intention information is difficult to interpret concretely. We have devised two encoders to model past observations and their corresponding estimated destinations. To simultaneously learn the coordinate information (x, y) from the past trajectory, the locations are reshaped into $N \times T_{past} \cdot 2$ and then encoded by a learnable linear transformation to obtain embedded past features E_{obs}. The transformer-encoder layer is then utilized to further explore the context information E_{con} of embedded past features E_{obs}. Residual connection [He *et al.*, 2016] incorporates above two hidden states and forms observation feature f_{obs}.

$$f_{obs} = \text{Encoder}_{obs}(\mathbf{X}). \tag{8}$$

The destination points are encoded by an MLP to get the destination feature f_{des} in Eq.(9), which is regarded as an implicit intention. The destination features of each independent agent are fused by cross combination to simulate the interaction between diverse destinations, as the intention of an agent may also be influenced by surrounding neighbors.

$$f_{des} = \text{MLP}_{des}(\widehat{\mathbf{\Omega}}). \tag{9}$$

In fact, f_{obs} and f_{des} are two independent variables, and they fail to reveal the influence of intrinsic intention. Additionally, we leverage the cross-attention mechanism to expose the intention feature f_{int}, setting the destination feature as query Q and the observation feature as key K. Thus, the intention feature can be calculated as follows:

$$f_{int} = \text{softmax}\left(\left(W_Q f_{des}\right)\left(W_K f_{obs}\right)^T\right) W_V f_{obs}, \tag{10}$$

where W_Q, W_K and W_V are three learnable weight matrices, the dimension of the former matrix is set the same as the destination feature f_{des}, and the dimension of the last two matrices are set the same as the observation feature f_{obs}.

2.5 Trajectory Inference

Noised-distribution estimation. The standard diffusion model has demonstrated a solid ability to learn sophisticated distributions of human trajectory [Gu *et al.*, 2022], but leading time-consuming denoising steps to generate high-quality samples. Inspired by LED [Mao *et al.*, 2023], which proposed an initialized diffusion model to accelerate the inference speed, we build a new noised trajectory distribution as the initiation to accelerate the denoising process. Our method can capture humans' intrinsic and extrinsic motivations to estimate high-quality trajectories. We concatenate the intrinsic intention feature and the extrinsic interaction feature to form the multi-modality fusion feature f_{fusion}, and set three learnable decoding modules to estimate the mean $\widetilde{\mu}_\theta$, standard deviation $\widetilde{\sigma}_\theta$ and normalized positions $\widetilde{\mathcal{P}}_\theta$. K samples are generated as follows:

$$f_{fusion} = f_{int} \oplus f_{ext}, \tag{11}$$

$$\widetilde{\mu}_\theta = \text{Decoder}_\mu(f_{fusion}) \in \mathbb{R}^{T_{pred} \times 2}, \tag{12}$$

$$\widetilde{\sigma}_\theta = \text{Decoder}_\sigma(f_{fusion}) \in \mathbb{R}, \tag{13}$$

$$\widetilde{\mathcal{P}}_\theta = \{\widetilde{p}_\theta^k | \widetilde{p}_\theta^k = \text{Decoder}_\mathcal{P}(f_{fusion}), k \in 1, ..., K\} \in \mathbb{R}^{K \times T_{pred} \times 2}, \tag{14}$$

where MLP is selected as the decoder for small parameters and computational complexity, k indicates the number of samples for each prediction. To avoid learning a raw sophisticated distribution as mentioned in [Mao *et al.*, 2023], reparameterization in Eq.(15) can be utilized to refine predicted trajectories.

$$\widetilde{Y}^k = \widetilde{\mu}_\theta + \widetilde{\sigma}_\theta \cdot \widetilde{\mathcal{P}}_\theta^k \in \mathbb{R}^{T_{pred} \times 2}. \tag{15}$$

Denoising procedure. Once we obtain the noised trajectories $\widetilde{\mathbf{Y}}$, we utilize an attention-based module to encode the context of observations $\mathbf{X}$ and estimate the noise to reduce by a series of learnable linear-transformation operations. The recovered trajectories are presented as the following formula:

$$\beta = 1 - \alpha, \tag{16}$$

$$\widehat{Y}^{\tau-1} = \frac{1}{\sqrt{\alpha^{\tau-1}}}\left(\widehat{Y}^\tau - \frac{\beta^{\tau-1}}{\sqrt{1-\bar{\alpha}^{\tau-1}}}\epsilon_\theta^{\tau-1}(\widetilde{Y}^\tau, \mathbf{X}, C, \tau)\right) + \sqrt{\beta^{\tau-1}}\mathbf{z} \in \mathbb{R}^{N \times K \times T_{pred} \times 2}, \tag{17}$$

where $\alpha^{\tau-1}$ and $\bar{\alpha}^{\tau-1} = \prod_{i=1}^{\tau-1} \alpha^i$ are parameters in standard diffusion model, β is a constant calculated by α, and $\mathbf{z}$ samples from the standard Gaussian distribution $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. $\epsilon_\theta^{\tau-1}$ is the noise estimated from noised trajectories $\widetilde{Y}^\tau$ with context embedding C and step τ.

2.6 Training Constraint

We design an end-to-end training strategy that collaboratively optimizes the intrinsic intention and extrinsic interaction feature encoders, including the denoising module. In LED [Mao *et al.*, 2023], the authors divide the training process into two stages and leverage the pre-trained denoising module to train the leapfrog initializer in the second stage. Unlike LED, we believe encoding multi-modality features to generate noised trajectory distribution will continuously affect denoising. To maintain a balance between the mentioned two processes, we optimize the Kullback-Leibler Divergence (D_{KL}) between $\widetilde{\mu}_\theta$ and μ_θ by calculating:

$$D_{KL} = \mathbb{E}_q\left[\lambda \|\widetilde{\mu}_\theta - \mu_\theta(Y^\tau, Y^0)\|^2\right] + C. \tag{18}$$

Moreover, we expect each sample to be close to the ground truth and leverage the marginal loss proposed in [Gupta *et al.*, 2018]. Our method generates multiple predicted trajectories by stochastic sampling from the estimated distribution $\mathcal{N}(\widetilde{\mu}_\theta, \widetilde{\sigma}_\theta \mathbf{I})$. The loss function employs $L2$ norm to minimize the distance between the prediction and the ground truth.

$$\mathcal{L}_{marginal} = \sum\nolimits_n^N \min_k^K \|Y_n - \widehat{Y}_n^k\|_2. \tag{19}$$

Inspired by [Weng *et al.*, 2023], we introduce a variation of $\mathcal{L}_{marginal}$ for joint training by swapping the order of taking the minimum over samples k and taking the accumulation over agents n.

$$\mathcal{L}_{joint} = \min_k^K \sum\nolimits_n^N \|Y_n - \widehat{Y}_n^k\|_2. \tag{20}$$

Additionally, to train the destination decoder, we set $\mathcal{L}_{des}$ with $L2$ norm as follows:

$$\mathcal{L}_{des} = \sum\nolimits_n^N \min_k^K \|\Omega_n - \widehat{\Omega}_n^k\|_2. \tag{21}$$

Scenes	Social-GAN	SoPhie	PECNet	Trajectron++	NMMP	MemoNet	GroupNet	MID	LED	**Ours**
ETH	0.87/1.62	0.70/1.43	0.54/0.87	0.61/1.02	0.61/1.08	0.40/0.61	0.46/0.73	0.39/0.66	0.39/0.58	**0.36/0.54**
HOTEL	0.67/1.37	0.76/1.67	0.18/0.24	0.19/0.28	0.33/0.63	0.11/0.17	0.15/0.25	0.13/0.22	0.11/0.17	**0.09/0.11**
UNIV	0.76/1.52	0.54/1.24	0.35/0.60	0.30/0.54	0.52/1.11	0.24/0.43	0.26/0.49	0.22/0.45	0.26/0.43	**0.16/0.25**
ZARA1	0.35/0.68	0.30/0.63	0.22/0.39	0.24/0.42	0.32/0.66	0.18/0.32	0.21/0.39	0.17/0.30	0.18/0.26	**0.13/0.20**
ZARA2	0.42/0.84	0.38/0.78	0.17/0.30	0.18/0.32	0.43/0.85	0.14/0.24	0.17/0.33	0.13/0.27	0.13/0.22	**0.11/0.17**
AVG	0.61/1.21	0.54/1.15	0.29/0.48	0.30/0.51	0.21/0.35	0.21/0.35	0.25/0.44	0.21/0.38	0.21/0.33	**0.17/0.25**

Table 1: Quantitative results of previously proposed methods compared with ours across dataset ETH-UCY. Two metrics of minADE_{20} / minFDE_{20} are reported for future 12 timesteps (4.8s). **Bold**/underlined fonts represent the best/second-best result, and the lower the better.

Time	Social-GAN	Trajectron++	SoPhie	NMMP	PECNet	MID	GroupNet	MemoNet	LED	**Ours**
4.8s	27.23/41.44	19.30/32.70	16.27/29.38	14.67/26.72	9.96/15.88	9.73/15.32	9.31/16.11	8.56/12.66	8.26/13.55	**7.83/12.43**

Table 2: Quantitative results of previously proposed methods compared with ours across dataset SDD. Two metrics of minADE_{20} and minFDE_{20} are reported for future 12 timesteps (4.8s). **Bold**/underlined fonts represent the best/second-best result, and the lower the better.

Sample consistency loss. The datasets of pedestrian trajectories are obtained from the specific scenes in reality. For socially acceptable prediction, the model needs to learn the sophisticated indeterminacy of trajectory distribution from the effect of intrinsic intention and extrinsic interaction.

The common multi-modal trajectory prediction approaches encourage the model to generate k potential samples. However, the consistency among each sample is not constrained, resulting in significant deviations between various samples generated under the same distribution. Increasing k is a solution to address large deviations [Gupta *et al.*, 2018], but it raises computational complexity. Empirically, as the number of k gradually increases, the model generates samples as much as possible, so an excellent trajectory close to the ground truth will be identified in this pile of samples with a high probability. Essentially, the sampling program can be regarded as a process of entropy increment. To emphasize the consistency of multi-modal prediction, we propose *sample consistency loss* to constrain the deviations among different k samples in the trajectory inference phase for each agent. Training our framework combined with $\mathcal{L}_{sc}$ improves the model's fitting ability without additional computation cost.

$$\mathcal{L}_{sc} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \| \sqrt{(Y_n - \mu)^2} - \sqrt{(\widehat{Y}_n - \widetilde{\mu})^2} \|_2, \tag{22}$$

where μ and $\widetilde{\mu}$ are the mean values of the ground truth $\mathbf{Y}$ and predicted trajectories $\widehat{\mathbf{Y}}$, respectively. See more details about how *sample consistency loss* will affect trajectory prediction in Sec. 3.5.

Total loss. To form our training loss, we compose each part together and set different weight coefficients λ as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{marginal} + \lambda_2 \mathcal{L}_{joint} + \mathcal{L}_{sc} + \mathcal{L}_{des} + D_{KL}. \tag{23}$$

3 Experiments

3.1 Datasets

The framework is trained and evaluated on two public benchmarks, including ETH-UCY [Pellegrini *et al.*, 2010], and Stanford Drone Dataset [Robicquet *et al.*, 2016].

ETH-UCY. Each dataset contains several different scenes, such as ETH and HOTEL in ETH, ZARA1, ZARA2 and UNIV in UCY. These subsets map from the real-world scenarios, including individual behaviors and group behaviors. The datasets are captured at 25 fps and annotated at 2.5 fps, and we use 3.2s (8 frames) as past input data to predict 4.8s (12 frames) future trajectories.

Stanford Drone Dataset (SDD). The dataset captures extensive human behaviors in campus from the bird's eye view. It covers no less than 40,000 human-scene interactions and 185,000 human-human interactions. Following previous works [Mangalam *et al.*, 2020; Xu *et al.*, 2022b], we use the standard train-test split and predict the future 4.8s (12 frames) using 3.2s (8 frames) past.

3.2 Implementation Details

In the intrinsic intention encoder, we initially use MLPs to learn the linear transformation of past observations and set the output dimension 256. We subsequently utilize a two-layer transformer-based module with head number 4 to encode past embeddings. In the extrinsic interaction encoder, the LSTM, in which the hidden dimension is set to 256, encodes temporal interaction features, and two graph attention layers with 4 heads are utilized to explore spatial interaction features. Then, the concatenate operation fuses a multi-modality feature with a dimension of 512. The decoders used MLPs to estimate the mean, standard deviation, and normalized positions with the same input dimension 1024. The number of samples K is set to 20 to generate multiple results. The model is trained by an AdamW optimizer [Loshchilov and Hutter, 2017] with a batch size 16 and an initial learning rate of 1×10^{-3} on a single 3090 GPU for 100 epochs. For the loss coefficient weights, we set $\lambda_1 = 5.0$ and $\lambda_2 = 1.0$.

3.3 Evaluation Protocol

We adopt two widely reported metrics to evaluate the prediction performance: the Average Displacement Error (ADE)

Intention f_{int}	Interaction f_{ext}	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
✓	-	0.46/0.74	0.14/0.20	0.24/0.41	0.20/0.36	0.15/0.24	0.24/0.39
-	✓	0.41/0.66	0.12/0.18	0.22/0.39	0.19/0.32	0.14/0.24	0.22/0.36
✓	✓	**0.35/0.50**	**0.11/0.14**	**0.20/0.31**	**0.17/0.26**	**0.12/0.19**	**0.19/0.28**

(a) Ablation experiments of the intrinsic intention and extrinsic interaction to trajectory prediction. minADE$_{20}$ / minFDE$_{20}$ (meters) are reported.

Loss $\mathcal{L}_m$	Loss $\mathcal{L}_j$	Avg Metric ADE↓	FDE↓	JADE↓	JFDE↓
✓	-	0.19	0.30	0.43	0.88
-	✓	0.33	0.53	0.34	0.59
✓	✓	**0.17**	**0.25**	**0.32**	**0.62**

(b) Ablations to study the influence of marginal metrics and joint metrics with the supervision of corresponding loss $\mathcal{L}_m$ and $\mathcal{L}_j$.

Table 3: Ablation experiments.

Feature f_{int}	Feature f_{ext}	Loss $\mathcal{L}_m$	Loss $\mathcal{L}_j$	Loss $\mathcal{L}_{sc}$	N-Sample K	Avg Metric ADE↓	FDE↓	JADE↓	JFDE↓
✓	-	✓	✓	-	20	0.33	0.48	0.55	1.06
✓	-	✓	✓	✓	20	0.30	0.40	0.42	0.85
-	✓	✓	✓	-	20	0.21	0.34	0.42	0.82
-	✓	✓	✓	✓	20	0.19	0.30	0.40	0.79
✓	✓	✓	✓	-	20	0.19	0.31	0.39	0.78
✓	✓	✓	✓	✓	20	**0.17**	**0.25**	**0.32**	**0.62**

Table 4: Ablations to study the influence of sample consistency loss $\mathcal{L}_{sc}$. To comprehensively analyze how this constraint affects multiple-prediction performance, we include other ablative parts. Average ADE/FDE and JADE/JFDE are reported.

and the Final Displacement Error (FDE) [Pellegrini *et al.*, 2009]. ADE measures the accuracy between the ground truth and predicted trajectory over all time steps. FDE measures the accuracy only at the predicted last point of the trajectory. However, a recent work [Weng *et al.*, 2023] thoughtfully analyzes only marginal metrics (ADE and FDE) during evaluation, which will result in an overly optimistic estimation of performance, and the author proposes joint metrics (JADE and JFDE). We also include the joint loss term as Eq.(20) and report evaluation results of joint metrics in our experiments. See more details in the supplementary material.

3.4 Quantitative Results

In order to further present the effectiveness of our proposed method, experiments have been performed to compare with the previous state-of-the-art methods across different datasets (e.g., Social-GAN [Gupta *et al.*, 2018], SoPhie [Sadeghian *et al.*, 2019], PECNet [Mangalam *et al.*, 2020], Trajectron++ [Salzmann *et al.*, 2020], NMMP [Hu *et al.*, 2020], MemoNet [Xu *et al.*, 2022b], GroupNet [Xu *et al.*, 2022a], MID [Gu *et al.*, 2022] and LED [Mao *et al.*, 2023]). Note that the results of LED are reproduced by us for fair comparison. We report two evaluation metrics, minADE$_K$ and minFDE$_K$ obtained from the Best-of-N [Gupta *et al.*, 2018], which selects the closest sample from K generations to the ground truth and calculates the ADE/FDE metric on it.

ETH-UCY. Table 1 shows the evaluation results of our method compared with 9 previous prominent methods. We can drop that DifTraj outperforms those previously proposed models by observing the average values of metrics ADE and FDE, and achieves the state-of-the-art performance. Specifically, compared with the current state-of-the-art method LED, our method reduces ADE from 0.21 to 0.17 and FDE from 0.33 to 0.25, achieving a 19.0% and a 24.2% improvement, respectively. An incredible phenomenon is that our method achieves the best performance among all subsets.

SDD. Table 2 shows the evaluation results on SDD datasets of our method compared with 9 previous prominent methods. We set a limitation of the sample number $K = 20$ spreading around all methods. The results indicate that our method can achieve a superior reduction on ADE of 7.83 and FDE of 12.43, which outperforms LED with a 5.2% improvement on ADE and a 8.3% improvement on FDE.

3.5 Ablation Studies

Ablative experiments are all conducted on ETH-UCY shown in Table 3 and Table 4, results isolate the contribution of each part to the prediction performance. Note that the ablation study is not affected by the sample number K, we set default $K = 20$.

Effect of the intrinsic intention and extrinsic interaction. We conduct ablation studies to analyze how our method predicts more precise and socially-acceptable trajectories by learning from the intrinsic intention and extrinsic interaction as shown in Table 3a. Note that we only change the input feature types with the same dimension and keep other settings consistent. On the one hand, leveraging the past observations and destinations to encode the intrinsic intention features, the estimated trajectory distributions are seriously disturbed by noises with a high indeterminacy, which confuses the denoising process. The results on five subsets of ETH-UCY are poor. In contrast, by utilizing the extrinsic interaction features to estimate pseudo trajectories, we see that the results after the denoising process achieve good performances on all five subsets. We can drop that intrinsic intention and extrinsic interaction are significant to trajectory prediction.

Effect of training with joint metrics. As the author advocated for the use of joint metrics over marginal metrics in [Weng *et al.*, 2023], we conduct ablations to explicitly study the influence of marginal metrics and joint metrics with the supervision of $\mathcal{L}_{marginal}$ and/or $\mathcal{L}_{joint}$ to train the model. Note that we utilize both the intrinsic intention and extrinsic interaction features. Intuitively, the joint loss focuses on multiple agent prediction, while the marginal loss cares for individual agent prediction. An interesting result in Table 3b is that the training incorporated with marginal loss $\mathcal{L}_{marginal}$ and joint loss $\mathcal{L}_{joint}$ leads to an improvement on

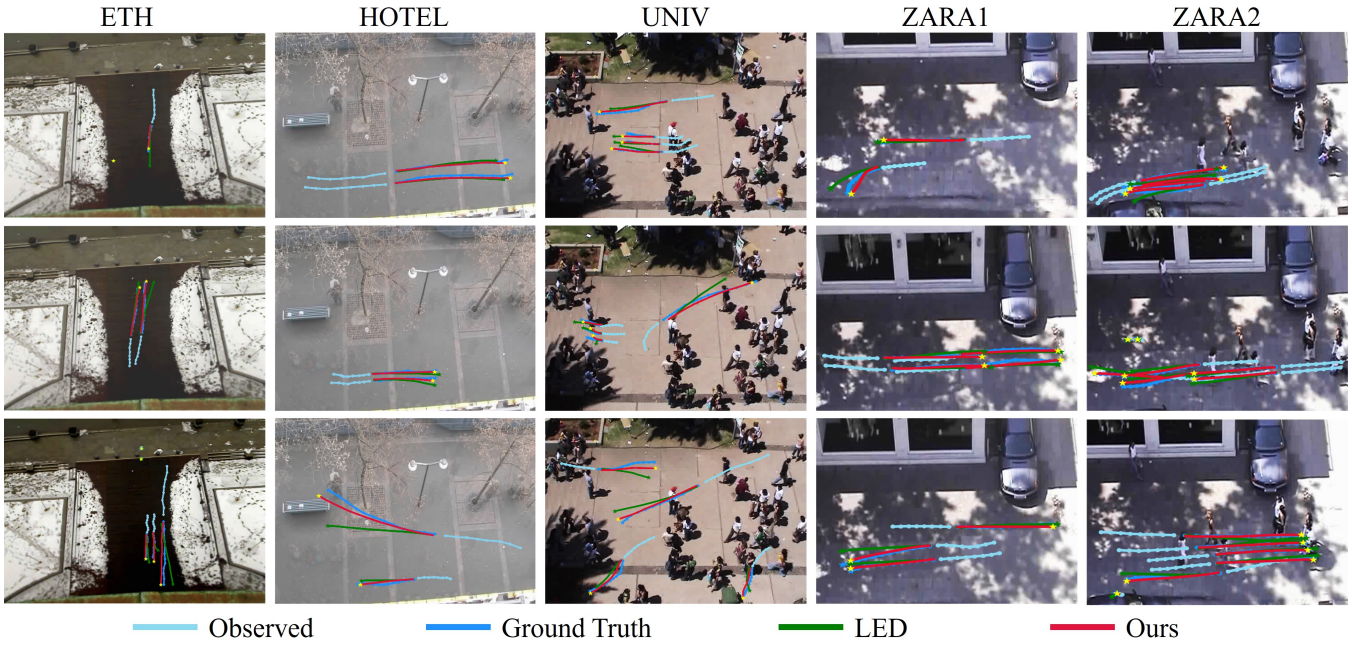

Figure 3: Visualization results on ETH-UCY dataset. Given the past observations (sky blue), we visualize the best predicted trajectory from 20 samples and destinations (yellow stars) by LED (green) and ours (red), the ground truth of trajectory is presented by blue line. We see that our predicted trajectory is more close to the ground truth than LED.

both ADE/FDE and JADE/JFDE metrics. A reasonable explanation is that training with marginal and joint loss encourages the model to produce accurate predictions for each agent in any of its K samples and to predict realistic trajectories from multi-agent interaction. The patterns of future human behavior obtain complete preservation. We have attempted to prove that the latter is more challenging to converge than the former. See more details in the supplementary material.

Effect of the sample consistency loss. To comprehensively analyze how this constraint affects multiple-prediction performance, we perform an ablation experiment and include other ablative parts shown in Table 4. As we have indicated that training with marginal loss and joint loss improves the performance in the last ablation experiment, we set them as default settings in this experiment. We see that i) only inputting intention features with the supervision of *sample consistency loss* reduces ADE/FDE from 0.33/0.48 to 0.30/0.40, only inputting interaction features with the supervision of *sample consistency loss* reduces ADE/FDE from 0.21/0.34 to 0.19/0.30, and inputting both of them with the supervision of *sample consistency loss* reduces ADE/FDE from 0.19/0.31 to 0.17/0.25; ii) JADE and JFDE are both improved under conditions of $\mathcal{L}_{sc}$ supervision; and iii) our novel loss helps model learn the sophisticated indeterminacy of trajectory distribution from the intrinsic intention and extrinsic interaction, and improves the prediction performance.

3.6 Qualitative Results

To have a more intuitive understanding of the performance of our method, we visualize the best-predicted trajectory from 20 samples and compare ours with LED in Figure 3. The results show that ours and LED fit the ground truth well, but the latter performs unsatisfactorily in the subset of UNIV with extensive interactive behaviors. Our method produces more accurate and socially-acceptable predictions that are close to humans' real intentions.

4 Conclusion

In this paper, we analyze the conditions of multi-modal trajectory prediction from two objective perspectives and propose a novel end-to-end framework based on the diffusion model to predict more precise and socially-acceptable trajectories for humans. Firstly, a spatial-temporal aggression module is built to extract the extrinsic interaction features for capturing socially-acceptable behaviors. Secondly, we explicitly construct the intrinsic intention module to obtain intention features for precise prediction. Finally, we estimate a noise trajectory distribution with these two features as the initiation of the diffusion model and leverage the denoising process to obtain the final trajectories. Furthermore, to reduce the noise of the initiative trajectory estimation, we present a novel *sample consistency loss* to constrain multiple predictions. Extensive results indicate that our method has achieved state-of-the-art performance on ETH-UCY and SDD datasets. In the future, we will try to extend our method to other sequence prediction tasks, such as visual tracking and video segmentation [Shen *et al.*, 2021; Dong *et al.*, 2023; Wu *et al.*, 2023; Wu *et al.*, 2022].

Acknowledgments

This work is partially funded by the Fundamental Research Funds for the Central Universities (Ref. No.: 2042022kf1198), the Science and Technology Development Fund (SKL-IOTSC(UM)-2024-2026) and the State Key Laboratory of Internet of Things for Smart City (University of Macau) (Ref. No.: SKL-IoTSC(UM)-2024-2026/ORP/GA04/2023), and Key Research and Development Project of Hubei Province (2022BAD175, 2022BCA009).

References

[Bae *et al.*, 2022] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Non-probability sampling network for stochastic human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6477–6487, 2022.

[Cheng *et al.*, 2020] Yujiao Cheng, Liting Sun, Changliu Liu, and Masayoshi Tomizuka. Towards efficient human-robot collaboration with robust plan recognition and trajectory prediction. *IEEE Robotics and Automation Letters*, 5(2):2602–2609, 2020.

[Dong *et al.*, 2023] Xingping Dong, Jianbing Shen, Fatih Porikli, Jiebo Luo, and Ling Shao. Adaptive siamese tracking with a compact latent network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8049–8062, 2023.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[Gu *et al.*, 2022] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17113–17122, 2022.

[Gupta *et al.*, 2018] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[Hu *et al.*, 2020] Yue Hu, Siheng Chen, Ya Zhang, and Xiao Gu. Collaborative motion prediction via neural motion message passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6319–6328, 2020.

[Kong *et al.*, 2020] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

[Makansi *et al.*, 2022] Osama Makansi, Julius von Kügelgen, Francesco Locatello, Peter Gehler, Dominik Janzing, Thomas Brox, and Bernhard Scholkopf. You mostly walk alone: Analyzing feature attribution in trajectory prediction. In *International Conference on Learning Representations*, 2022.

[Mangalam *et al.*, 2020] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European Conference on Computer Vision*, pages 759–776, 2020.

[Mangalam *et al.*, 2021] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15233–15242, 2021.

[Mao *et al.*, 2023] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5517–5526, 2023.

[Nichol and Dhariwal, 2021] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171, 2021.

[Pellegrini *et al.*, 2009] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 261–268, 2009.

[Pellegrini *et al.*, 2010] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European Conference on Computer Vision*, pages 452–465, 2010.

[Qiao *et al.*, 2014] Shaojie Qiao, Dayong Shen, Xiaoteng Wang, Nan Han, and William Zhu. A self-adaptive parameter selection trajectory prediction approach via hidden markov models. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):284–296, 2014.

[Robicquet *et al.*, 2016] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European Conference on Computer Vision*, pages 549–565, 2016.

[Sadeghian *et al.*, 2019] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. 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*, pages 1349–1358, 2019.

[Salzmann *et al.*, 2020] Tim Salzmann, B. Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pages 683–700, 2020.

[Shen *et al.*, 2021] Jianbing Shen, Yuanpei Liu, Xingping Dong, Xiankai Lu, Fahad Shahbaz Khan, and Steven Hoi. Distilled siamese networks for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8896–8909, 2021.

[Shu *et al.*, 2018] Xiangbo Shu, Jinhui Tang, Guo-Jun Qi, W. Liu, and Jian Yang. Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:1110–1118, 2018.

[Sohn *et al.*, 2015] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

[Su *et al.*, 2017] Hang Su, Jun Zhu, Yinpeng Dong, and Bo Zhang. Forecast the plausible paths in crowd scenes. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2772–2778, 2017.

[Teeti *et al.*, 2022] Izzeddin Teeti, Salman Khan, Ajmal Shahbaz, Andrew Bradley, and Fabio Cuzzolin. Vision-based intention and trajectory prediction in autonomous vehicles: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5630–5637, 2022.

[Tran *et al.*, 2021] Hung Tran, Vuong Le, and Truyen Tran. Goal-driven long-term trajectory prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 796–805, 2021.

[Valera and Velastin, 2005] Maria Valera and Sergio A Velastin. Intelligent distributed surveillance systems: a review. *IEE Proceedings-Vision, Image and Signal Processing*, 152(2):192–204, 2005.

[Veličković *et al.*, 2017] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[Weng *et al.*, 2023] Erica Weng, Hana Hoshino, Deva Ramanan, and Kris Kitani. Joint metrics matter: A better standard for trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20315–20326, 2023.

[Wong *et al.*, 2022] Conghao Wong, Beihao Xia, Ziming Hong, Qinmu Peng, Wei Yuan, Qiong Cao, Yibo Yang, and Xinge You. View vertically: A hierarchical network for trajectory prediction via fourier spectrums. In *European Conference on Computer Vision*, pages 682–700, 2022.

[Wu *et al.*, 2022] Dongming Wu, Xingping Dong, Ling Shao, and Jianbing Shen. Multi-level representation learning with semantic alignment for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4996–5005, 2022.

[Wu *et al.*, 2023] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14633–14642, 2023.

[Xu *et al.*, 2018] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and M. Kankanhalli. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia*, 22:1423–1432, 2018.

[Xu *et al.*, 2022a] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6507, 2022.

[Xu *et al.*, 2022b] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2022.

[Yuan *et al.*, 2021] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021.