SocialCVAE: Predicting Pedestrian Trajectory via Interaction Conditioned Latents

Anonymous submission

Abstract

Pedestrian trajectory prediction is the key technology in many applications for providing insights into human behavior and anticipating human future motions. Most existing empirical models are explicitly formulated by observed human behaviors using explicable mathematical terms with deterministic nature, while recent work has focused on developing hybrid models combined with learning-based techniques for powerful expressiveness while maintaining explainability. However, the deterministic nature of the learned steering behaviors from the empirical models limits the models' practical performance. To address this issue, this work proposes the social conditional variational autoencoder (SocialCVAE) for predicting pedestrian trajectories, which employs a CVAE to explore behavioral uncertainty in human motion decisions. SocialCVAE learns socially reasonable motion randomness by utilizing a socially explainable interaction energy map as the CVAE's condition, which illustrates the future occupancy of each pedestrian's local neighborhood area. The energy map is generated using an energy-based interaction model, which anticipates the energy cost (i.e., repulsion intensity) of pedestrians' interactions with neighbors. Experimental results on two public benchmarks including 25 scenes demonstrate that SocialCVAE significantly improves prediction accuracy compared with the state-of-the-art methods, with up to 16.85% improvement in Average Displacement Error (ADE) and 69.18% improvement in Final Displacement Error (FDE). The code will be released upon acceptance.

1 Introduction

Pedestrian trajectory prediction is a vital task in intelligent systems for understanding human behavior and anticipating future motions. Predicting the future movements of pedestrians in complex environments is challenging due to the highly dynamic and subtle nature of human interactions. Empirical methods explicitly model interactions for crowd motion prediction, e.g., rule-based model (Reynolds 1987; Reynolds et al. 1999), force-based model (Helbing and Molnar 1995; Karamouzas, Skinner, and Guy 2014) and energy-based model (Guy et al. 2010; Karamouzas et al. 2017). These models are explainable but with lower predictive accuracy, as they cannot fit observed data precisely. In contrast, various methods based on deep neural nets have been proposed with social interaction modeling by employing social pooling mechanism (Alahi et al. 2016; Gupta et al. 2018), graph-based modeling (Mohamed et al. 2020; Bae and Jeon 2021), and attention mechanism (Mangalam et al. 2020; Shi et al. 2021). While they achieve expressive power and generalization ability, their black-box nature makes the learned model less interpretable to human understanding. It remains a challenge to explore the trade-off between model explainability and prediction capability. Recent research effort has been focused on exploring the aforementioned trade-off by designing hybrid models that combine deep neural nets with explainable interaction (Kothari, Sifringer, and Alahi 2021; Yue, Manocha, and Wang 2022). However, their prediction accuracy suffers from the deterministic nature of the physics-driven behaviors (Yue, Monocha, and Wang 2023).

To overcome the challenges while retaining the advantages of hybrid methods, we propose SocialCVAE, a new hybrid model for pedestrian trajectory prediction that combines an energy-based interaction model for socially explainable interaction anticipations with an interaction-conditioned CVAE for multimodal prediction. Fig. 1 illustrates the framework of our method. SocialCVAE takes advantage of the data-driven optimization model (Xiang et al. 2023) to quantify the interaction energy cost (i.e., repulsion intensity) of the temporal coarse predictions and explicitly represent the interaction energies into the local energy map. Using the CVAE model conditioned on the interaction energy map, SocialCVAE learns socially reasonable residuals for the temporal motion decisions. Similar to the previous methods (Zhou et al. 2021; Yue, Manocha, and Wang 2022) that achieve state-of-the-art (SOTA) performance, we employ the recursive prediction scheme to update future trajectories step by step with the input trajectories at each step including the updated trajectories.

We conduct extensive experiments on two popular benchmark datasets (ETH-UCY (Pellegrini et al. 2009; Lerner, Chrysanthou, and Lischinski 2007) and SDD (Robicquet et al. 2016)), and demonstrate SocialCVAE's superiority over existing state-of-the-art methods in terms of prediction accuracy. Furthermore, our results highlight the effectiveness of using an energy-based interaction model for pedestrian trajectory prediction and provide insights into how to better model pedestrian behavior in complex environments. The main contributions are concluded as follows:

- We propose a novel multimodal pedestrian trajectory prediction model (SocialCVAE) that leverages the advan-

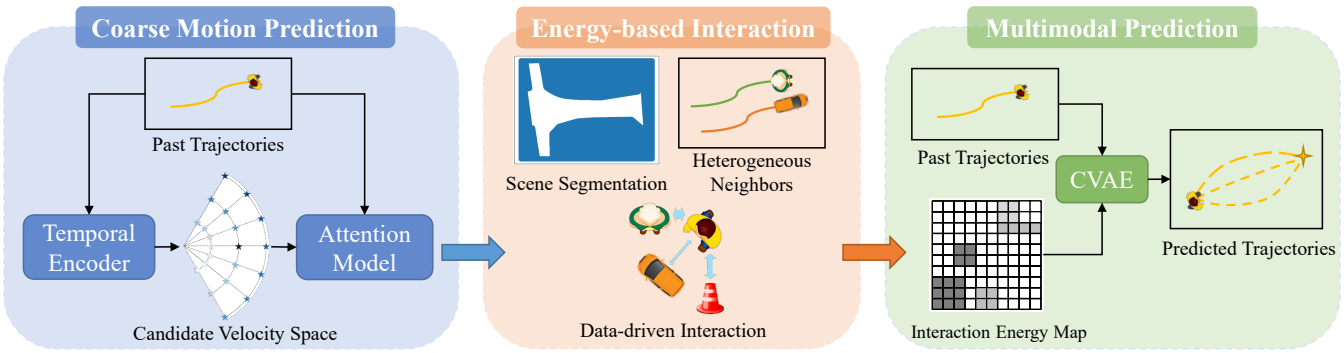


Figure 1: The framework of SocialCVAE. (a) The coarse motion prediction model learns the temporal motion tendencies and predicts a preferred new velocity for each pedestrian. (b) The energy-based interaction model constructs a local interaction energy map to anticipate the cost of pedestrian interactions with heterogeneous neighbors, including pedestrians, static environmental obstacles found in the scene segmentation (e.g., buildings), and dynamic environmental obstacles (e.g., vehicles). (c) The multimodal prediction model predicts future trajectories using a CVAE model conditioning on the past trajectories and the interaction energy map.

tages of both empirical and learning-based approaches for better prediction performance and interpretability of motion decisions.

- SocialCVAE explores the behavioral uncertainty of human motion by introducing socially explainable interaction energy maps generated from an energy-based interaction model. Both the quantitative and quality results of SocialCVAE demonstrate that the energy-based interaction helps the model better understand the social relationships between pedestrians, leading to improved prediction performance.

2 Related Works

Energy-based interaction methods. Considering the nonlinear nature of pedestrian motion dynamics that pedestrians try to anticipate and react to the future trajectories of their neighbors for collision avoidance (Karamouzas, Skinner, and Guy 2014), energy-based methods (Karamouzas et al. 2017; Ren et al. 2019; Xiang et al. 2023) predicts pedestrians' future trajectories by minimizing the anticipated social interaction cost calculated by energy functions, i.e., the anticipated repulsion intensity from neighbors. These models explicitly predict pedestrians' future motion by consuming minimum interaction cost, but provide less prediction accuracy due to relying solely on explicit motion features such as velocity. Comparatively, our method is a hybrid model that leverages the social explainability of energy-based interaction models with the prediction capability of deep-learning models, resulting in better prediction performance.

Data-driven methods. With advances in data acquisition techniques, deep learning methods have been proposed and achieved impressive results in predicting pedestrian trajectories. RNN structure has widely been used to capture temporal dependencies while considering social interactions using pooling mechanism (Alahi et al. 2016; Bisagno, Zhang, and Conci 2018; Gupta et al. 2018) or attention mechanism (Vemula, Muelling, and Oh 2018; Sadeghian et al. 2019; Salzmann et al. 2020; Xu, Hayet, and Karamouzas 2022). Graph-based models that utilize distance-based physical adjacency matrices (Mohamed et al. 2020; Bae and Jeon 2021; Xu et al. 2022) or attention-based learnable adjacency matrices (Huang et al. 2019; Shi et al. 2021; Duan et al. 2022; Wu et al. 2023) to learn pedestrian social interactions have also been developed. Besides, transformer-based models incorporate attention mechanisms (Yu et al. 2020; Yuan et al. 2021) to model social interaction for better performance in pedestrian trajectory prediction tasks.

Recently, prediction accuracy improvement has been made by NSP-SFM (Yue, Manocha, and Wang 2022), a multimodal prediction model which is a hybrid of steering behavior learning based on conservative position-dependent forces with unexplainable randomness learning. However, the deterministic force-driven behavior of NSP may result in performance degradation (Yue, Monocha, and Wang 2023). Different from NSP-SFM, our method combines the energy-based interaction model for explicit interaction cost anticipation with interaction-conditioned human motion uncertainty learning, resulting in providing socially reasonable randomness of future motion and yielding superior prediction performance.

3 Methodology

3.1 Problem Formulation

Pedestrian trajectory prediction aims to predict the positions of pedestrians' trajectories in a traffic scenario. Given the observed trajectories $\mathcal{X}_o = \{X_1, X_2, \ldots, X_n\}_{t=1}^{T_{obs}}$ of n pedestrians over T_{obs} time steps, where $X_i = \{\mathbf{x}_i^1, \ldots, \mathbf{x}_i^{T_{obs}}\}$ includes the observed spatial (2D-Cartesian) coordinates of pedestrian i, our goal is to predict the pedestrians' future trajectories $\hat{\mathcal{X}}_{pred}$ over the next T_{pred} steps. Regarding $\mathcal{X}_o$, the scene segmentation

S (see Fig. 1b) and the observed trajectories of other dynamically moving obstacles (e.g., vehicles) X_d as inputs, the prediction task is formulated as:

$$\hat{\mathcal{X}}_{pred} = f(\mathcal{X}_{\{o,d\}}, S), \tag{1}$$

where f is the model.

3.2 Framework

The framework of our method is illustrated in 1. Overall, SocialCVAE learns the uncertainty of human motion and implements it as predicting residuals of a coarse prediction. Specifically, a coarse prediction model (Fig. 1a) predicts a temporally reasonable preferred velocity and a new position for each pedestrian by aggregating the information from a discrete candidate velocity space, which is built based on the learned temporal tendency from an RNN-structured temporal encoder and includes possible temporal motion decisions (velocities). Then, an energy-based interaction model (Fig. 1b) anticipates the social interaction cost of the preferred velocity with heterogeneous neighbors, including interactions with pedestrians, static obstacles (e.g., buildings) obtained from the scene segmentation, and dynamic obstacles (e.g., vehicles). We map the interaction energy with the neighbors onto a local energy map to represent the future occupancy of the local neighborhood area. Finally, a CVAE model (Fig. 1c), which is conditioned on both past trajectories and the interaction energy map, predicts the socially reasonable residuals of the preferred new position and generates multimodal future trajectories.

3.3 Coarse Motion Prediction

The coarse motion prediction model predicts a temporally reasonable future motion for each pedestrian based on the trajectories in the past T_{obs} time steps.

Temporal motion tendency learning. We employ a recurrent neural network with one LSTM layer (Hochreiter and Schmidhuber 1997) to capture the temporal motion dependency and predict future motion. Given the hidden state h_i^t of each pedestrian i at time step t, a temporal extrapolation velocity $\bar{\mathbf{v}}_i^{t+1}$ can be obtained:

$$\begin{aligned} h_i^t &= \text{LSTM}(h_i^{t-1}, Relu(\phi(\mathbf{x}_i^t, \mathbf{v}_i^t))), \\ \bar{\mathbf{v}}_i^{t+1} &= \phi(h_i^t), \end{aligned} \tag{2}$$

where $\phi(\cdot)$ represents Linear transformation, $\mathbf{x}_i^t$ and $\mathbf{v}_i^t$ are the current location and velocity.

As human behavior is diverse and uncertain, multiple reasonable motion decisions exist for pedestrians. In our method, the possible motion decisions are explicitly modeled as the velocity candidates in a discrete candidate velocity space $V_{i,C}^t$, which is generated based on the temporal extrapolation velocity $\bar{\mathbf{v}}_i^{t+1}$. An illustration of $V_{i,C}^t$ is shown in Fig. 2. $V_{i,C}^t$ is a velocity set with size $4k_rk_\theta$.

Coarse Trajectory prediction. After obtaining the candidate velocity space representing multiple motion decisions, we need to optimize for the best one as the coarse motion

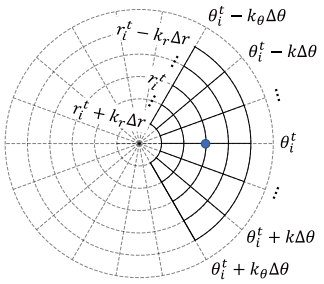


Figure 2: An illustration of the discrete candidate velocity space $V_{i,C}^t$ in a polar coordination system. The blue point represents the time extrapolation velocity $\bar{\mathbf{v}}_i^{t+1}$, with r_i^t and θ_i^t denoting the magnitude and angle of $\bar{\mathbf{v}}_i^{t+1}$. The polar space is discretized into a grid, with a predefined cell side length Δr and $\Delta\theta$ for the magnitude and angle axes. The velocity candidates in $V_{i,C}^t$ are represented by the intersection points of the solid lines, centered at $\bar{\mathbf{v}}_i^{t+1}$ within k_r grid cells on the magnitude axis and k_θ on the angle axis.

tendency for the subsequent time step. We adopt the attention mechanism (Vaswani et al. 2017) to score the relation between the trajectories $X_{i,P}^t$ in the past T_{obs} time steps and the velocity candidates from $V_{i,C}^t$, aggregate the information from $V_{i,C}^t$, and then obtain the future trajectory representation. The attention score matrix $\widetilde{A}_i^t$ is calculated like follows:

$$\begin{aligned} &F_{i,P}^t = MLP_1(X_{i,P}^t),\ F_{i,C}^t = MLP_2(\mathbf{V}_{i,C}^t), \\ &\widetilde{A}_i^t = \text{Softmax}\left(\frac{\phi(F_{i,P}^t)\phi(F_{i,C}^t)^T}{\sqrt{d_F}}\right), \end{aligned} \tag{3}$$

$\sqrt{d_F}$ is the scaled factor for ensuring numerical stability (Vaswani et al. 2017). Then the future trajectory representation F_i^t can be obtained:

$$F_i^t = F_{i,P}^t + \widetilde{A}_i^t F_{i,C}^t, \tag{4}$$

thus we can predict a preferred velocity $\widetilde{\mathbf{v}}_i^{t+1}$ for each pedestrian as a coarse motion decision, followed by predicting the coarse preferred new position $\widetilde{\mathbf{x}}_i^{t+1}$, that is

$$\begin{aligned} \widetilde{\mathbf{v}}_i^{t+1} &= MLP_3(F_i^t), \\ \widetilde{\mathbf{x}}_i^{t+1} &= \mathbf{x}_i^t + \widetilde{\mathbf{v}}_i^{t+1}\Delta t, \end{aligned} \tag{5}$$

where Δt is the horizon of a time step.

3.4 Energy-based Interaction Anticipating

As humans anticipate and react to the future trajectories of their neighbors for collision avoidance, we employ an energy-based interaction model similar to (Xiang et al. 2023) to calculate the interaction cost (i.e., repulsion intensity) driven by $\widetilde{\mathbf{v}}_i^{t+1}$. Our interaction model considers heterogeneous neighbors within a local neighborhood, which is a square area centering with the pedestrian (see Fig. 3a), including pedestrians, static obstacles (e.g., buildings), and dynamic obstacles (e.g., vehicles).

(a) (b)

Figure 3: An example of a pedestrian's square-shaped local interaction area. (a) The focal pedestrian at the center of the orange square interacts with heterogeneous neighbors within the square. (b) The interaction energy is recorded at the predicted local location of each neighbor (The darker color represents the higher value of interaction energy).

Interaction energy. Given the preferred velocity $\widetilde{\mathbf{v}}_i^{t+1}$ calculated by Eq. (5), the interaction energy is calculated based on the predicted distance $\widetilde{d}_{ij}^{t+1}$ of pedestrian i to the neighbor j at the next time step:

$$e_{ij}^t = e^{(1-\widetilde{d}_{ij}^{t+1}/d_s)}, \tag{6}$$

the neighbor j is assumed to hold its current velocity $\mathbf{v}_j^t$ for moving in the next time step. d_s is a hyperparameter as a scaling factor. Higher interaction energy means that the pedestrian is more likely to collide with the neighbor and vice versa. Considering the pedestrian may collide with the neighbor during a time step, $\widetilde{d}_{ij}^{t+1}$ is calculated as:

$$\widetilde{d}_{ij}^{t+1} = \|\mathbf{x}_{ij}^t + \widetilde{\mathbf{v}}_{ij}^{t+1} \cdot \widetilde{\tau}_{ij}^{t+1}\|_2, \tag{7}$$

where $\mathbf{x}_{ij}^t = \mathbf{x}_i^t - \mathbf{x}_j^t$, $\widetilde{\mathbf{v}}_{ij}^{t+1} = \widetilde{\mathbf{v}}_i^{t+1} - \mathbf{v}_j^t$. $0 \leq \widetilde{\tau}_{ij}^{t+1} \leq \Delta t$ is the predicted traveled time in the next time step, and it is calculated by clamping the solution of the following quadratic function:

$$\begin{aligned} \tau_{ij}^{t+1} &= \arg\min_\tau \|\mathbf{x}_{ij}^t + \widetilde{\mathbf{v}}_{ij}^{t+1} \cdot \tau\|_2, \\ \widetilde{\tau}_{ij}^{t+1} &= \text{clamp}(\tau_{ij}^{t+1}, [0, \Delta t]). \end{aligned} \tag{8}$$

Socially Explainable Energy Map. After calculating the interaction energies that anticipate and quantify the repulsion intensities from the neighbors, we project the interaction energies onto an energy map M_i^t, which has the same size as the local interaction area, in order to explicitly indicate the socially anticipated occupancy of each point within the local interaction area.

M_i^t is initialized as a zero matrix with size $L \times L$, where L is the side length of the local interaction area. A zero value in M_i^t means no occupancy, i.e., no risk of collision at this position during the next time step. The interaction energy e_{ij}^t calculated by Eq. (6) represents the occupancy of the neighbor j's future location $\widetilde{\mathbf{x}}_j^{t+1} = \mathbf{x}_j^t + \mathbf{v}_j^t \widetilde{\tau}_{ij}^{t+1}$. Fig. 3b illustrates the interaction energy map. Notably, to avoid performance degradation caused by the sparse matrix M_i^t, our method regards dynamic interaction neighbors as entities with a specific shape, and the points occupied inside the entity are allocated with the calculated interaction energy. A pedestrian neighbor is regarded as a square-shaped entity centered at $\widetilde{\mathbf{x}}_j^{t+1}$ with side length L_{ped}; for other types of dynamic neighbors (e.g., vehicles and bicycles), as we don't specify the accurate type of those neighbors, their occupied points are those inside the bounding box from the raw dataset. For the static neighbors (e.g., buildings) from the scene segmentation, the interaction neighbors are the points labeled impassable to pedestrians.

For a point $\mathbf{p} = (x_p, y_p)$ in the local interaction area, the corresponding value in the energy map is calculated as:

$$M_i^t[x_p, y_p] = \sum_{j \in \Omega(\mathbf{p})} w_{R(j)} \cdot e_{ij}^t, \tag{9}$$

where $\Omega(\mathbf{p})$ is the set of neighbors who are anticipated to occupy point $\mathbf{p}$, $R(j)$ is the type of neighbor $j \in \Omega(\mathbf{p})$, and $w_{R(j)}$ is the trainable weight of this neighbor type. The energy map contributes to the model for a better understanding of the future social relationship with the interaction neighbors.

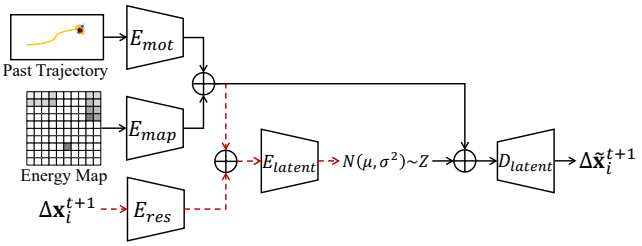


Figure 4: The architecture of the interaction-conditioned CVAE model. $\bigoplus$ represents the concatenation operation. Red dotted lines denote that the layers are only performed during training. All the components in the CVAE model are built using MLPs.

3.5 Multi-Modal Trajectory prediction.

To capture the uncertainty of human motion, we employ an interaction-conditioned CVAE model for multimodal trajectory forecasting. The architecture of our CVAE model is illustrated in Fig. 4. Different from the previous models that learn unexplainable randomness using CVAEs (Zhou et al. 2021; Yue, Manocha, and Wang 2022; Zhou et al. 2023), our model takes the pedestrian's past trajectories $X_{i,P}^t$ and the socially explainable energy map M_i^t as input to reconstruct the position residual $\Delta\mathbf{x}_i^{t+1} = \mathbf{x}_i^{t+1} - \widetilde{\mathbf{x}}_i^{t+1}$ between the ground-truth future position $\mathbf{x}_i^{t+1}$ and the predicted preferred position $\widetilde{\mathbf{x}}_i^{t+1}$ calculated by Eq. (5). As a result, the CVAE can learn socially reasonable randomness from data.

In the training process, the CVAE model firstly obtains the encodings of motion $F_{i,mot}^t$ and the encodings of the ground-truth position residual $F_{i,res}^{t+1}$:

$$\begin{aligned} F_{i,mot}^t &= E_{mot}(X_{i,P}^t) \oplus E_{map}(M_i^t), \\ F_{i,res}^{t+1} &= E_{res}(\Delta\mathbf{x}_i^{t+1}), \end{aligned} \tag{10}$$

where E_{mot}, E_{map} and E_{res} are the encoders. Then, a latent encoder E_{latent} generates the parameters (μ_i^t, σ_i^t) of a latent distribution:

$$(\mu_i^t, \sigma_i^t) = E_{latent}(F_{i,mot}^t \oplus F_{i,res}^{t+1}). \tag{11}$$

When predicting the future trajectory of a pedestrian, a latent decoder D_{latent} is employed to estimate a position residual $\Delta\widetilde{\mathbf{x}}_i^{t+1}$:

$$\Delta\widetilde{\mathbf{x}}_i^{t+1} = D_{latent}(F_{i,mot}^t \oplus Z_i^t), \tag{12}$$

where the latent variable Z_i^t is randomly sampled from a normal Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. Finally, the predicted position is:

$$\hat{\mathbf{x}}_i^{t+1} = \widetilde{\mathbf{x}}_i^{t+1} + \Delta\widetilde{\mathbf{x}}_i^{t+1}. \tag{13}$$

3.6 Loss Function

Our model is trained end-by-end by minimizing a multi-task loss:

$$\mathcal{L} = \frac{1}{nT_{pred}} \sum_{i=1}^{n} \sum_{t=T_{obs}+1}^{T_{end}} (\lambda_1 \mathcal{L}_{coarse} + \lambda_2 \mathcal{L}_{KL} + \lambda_3 \mathcal{L}_{pred}), \tag{14}$$

where $T_{end} = T_{obs} + T_{pred}$ is the last time step of the prediction time horizon, λ_1, λ_2 and λ_3 are the loss weights. $\mathcal{L}_{coarse}$ is the position loss for training the coarse prediction model, which measures the distance between each preferred new position with the ground truth. $\mathcal{L}_{KL}$ is the Kullback-Leibler (KL) divergence loss for training the CVAE model, which measures the distance between the sampling distribution of the latent variable learned at the training stage with the sampling normal Gaussian distribution at the test stage. $\mathcal{L}_{pred}$ is the predicted position loss for training the CVAE model, which measures the distance between each predicted position residual with the ground truth. That is:

$$\begin{aligned} \mathcal{L}_{coarse} &= \|\widetilde{\mathbf{x}}_i^t - \mathbf{x}_i^t\|_2, \\ \mathcal{L}_{\mathrm{KL}} &= D_{\mathrm{KL}}(\mathcal{N}(\mu_i^t, \sigma_i^t) || \mathcal{N}(0, \mathbf{I})), \\ \mathcal{L}_{pred} &= \|\Delta\widetilde{\mathbf{x}}_i^t - \Delta\mathbf{x}_i^t\|_2. \end{aligned} \tag{15}$$

4 Evaluation

4.1 Experiment Setup

Datasets. To evaluate the effectiveness of our method, we conduct extensive experiments on two widely used datasets in pedestrian trajectory prediction tasks: ETH-UCY dataset (Pellegrini et al. 2009; Lerner, Chrysanthou, and Lischinski 2007) and Stanford Drone Dataset (SDD) (Robicquet et al. 2016). ETH-UCY includes pedestrians' trajectories in 5 scenes (ETH, HOTEL UNIV, ZARA1, and ZARA2). We follow the leave-one-out strategy (Mangalam et al. 2021) for training and evaluation. SDD contains pedestrians' trajectories in 20 scenes. For SDD, we follow the data segmentation as (Yue, Manocha, and Wang 2022) for training and evaluation. Following the common practice (Mangalam et al. 2021; Yue, Manocha, and Wang 2022), the raw trajectories are segmented into 8-second trajectory segments with time step $\Delta t = 0.4s$, we train the model to predict the future $4.8s$ (12 frames) based on the observed $3.2s$ (8 frames).

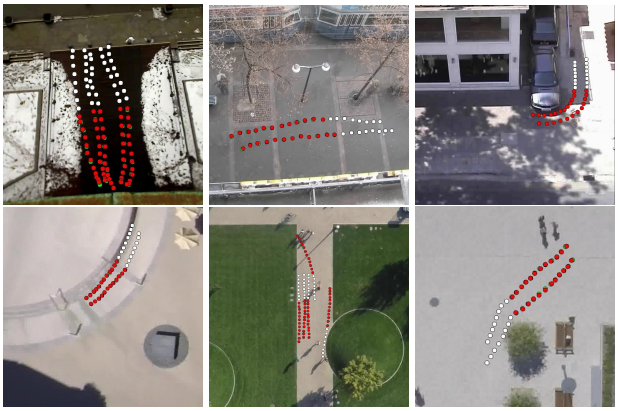


Figure 5: Visualization results of our method. The visualized trajectories are the best predictions sampled from 20 trials. The white, green, and red dots represent the observed, ground-truth, and predicted trajectories respectively.

Evaluation Metrics. We adopt the two widely used metrics, *Average Displacement Error* (ADE) and *Final Displacement Error* (FDE), to quantify the performance of our model. ADE computes the average L2 distance between the prediction and the ground truth over all predicted time steps. FDE calculates the L2 distance between the predicted final location and the ground-truth final location at the end of the prediction horizon. We follow the previously commonly used measurement to report the performances of the best of 20 predicted trajectories. Similar to (Zhou et al. 2021; Yue, Manocha, and Wang 2022; Zhou et al. 2023), we sample 20 future points at each prediction time step and select the best one as the predicted result.

Environment. Our model was implemented in PyTorch on a desktop computer running Ubuntu 20.04 containing an Intel ® Core™ i7 CPU and an NVIDIA GTX 3090 GPU. The model is trained end-to-end with an Adam optimizer with a learning rate 0.0001. We trained the ETH-UCY for 100 epochs and SDD for 150 epochs.

4.2 Quantitative Evaluation.

Quantitative Comparisons. We compare SocialCVAE with state-of-the-art models in recent years. The experimental results on $\mathrm{ADE}_{20}/\mathrm{FDE}_{20}$ are presented in Tab. 1 for ETH-UCY and Tab. 2 for SDD, showing that SocialCVAE achieves state-of-the-art performance on both datasets. Compared with the SOTA baseline methods, our method achieves performance improvement by 66.67% for FDE on ETH-UCY and 16.85%/69.18% for ADE/FDE on SDD. The main difference between SocialCVAE and the baseline methods is that our interaction-conditioned CVAE model learns socially reasonable motion randomness. The quantitative results demonstrate that SocialCVAE works well for better prediction performance.

Ablation study. We conduct ablative experiments to show the effectiveness of the key components in our model.

Table 1: Quantitative comparison with state-of-the-art methods on ETH-UCY for ADE_{20}/FDE_{20}. The bold/underlined font represents the best/second best result. The prediction results are measured in meters. Previous SOTA methods labeled by * also employ the recursive prediction scheme.

Model		ETH	Hotel	UNIV	ZARA1	ZARA2	AVG
S-GAN	(Gupta et al. 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
Sophie	(Sadeghian et al. 2019)	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.51/1.15
Trajectron++	(Salzmann et al. 2020)	0.39/0.83	0.12/0.21	0.20/0.44	0.15/0.33	0.11/0.25	0.19/0.41
PECNet	(Mangalam et al. 2020)	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48
YNET	(Mangalam et al. 2021)	0.28/0.33	0.10/0.16	0.24/0.41	0.17/0.27	0.13/0.22	0.18/0.27
Social-VAE	(Xu, Hayet, and Karamouzas 2022)	0.41/0.58	0.13/0.19	0.21/0.36	0.17/0.29	0.13/0.22	-
CAGN	(Duan et al. 2022)	0.41/0.65	0.13/0.23	0.32/0.54	0.21/0.38	0.16/0.33	0.25/0.43
SIT	(Shi et al. 2022)	0.39/0.61	0.13/0.22	0.29/0.49	0.19/0.31	0.15/0.29	0.23/0.38
MSRL	(Wu et al. 2023)	0.28/0.47	0.14/0.22	0.24/0.43	0.17/0.30	0.14/0.23	0.19/0.33
S-CSR*	(Zhou et al. 2021)	0.19/0.35	0.06/0.07	0.13/0.21	0.06/0.07	<u>0.05</u>/0.08	<u>0.10</u>/0.16
NSP-SFM*	(Yue, Manocha, and Wang 2022)	<u>0.07</u>/<u>0.09</u>	<u>0.03</u>/<u>0.07</u>	**0.03**/<u>0.04</u>	**0.02**/<u>0.04</u>	**0.02**/<u>0.04</u>	**0.03**/<u>0.06</u>
CSR*	(Zhou et al. 2023)	0.28/0.53	0.07/0.08	0.24/0.35	0.07/0.09	0.05/0.09	0.14/0.23
Ours*	-	**0.06/0.04**	**0.025/0.01**	**0.03/0.03**	**0.02/0.01**	**0.02/0.01**	**0.03/0.02**

Table 2: Quantitative comparison with state-of-the-art methods on SDD for ADE_{20}/FDE_{20}. The bold/underlined font represents the best/second best result. The prediction results are measured in pixels. Previous SOTA methods labeled by * also employ the recursive prediction scheme.

Model		ADE	FDE
S-GAN	(Gupta et al. 2018)	27.23	41.44
Sophie	(Sadeghian et al. 2019)	16.27	29.38
PECNet	(Mangalam et al. 2020)	9.96	15.88
YNET	(Mangalam et al. 2021)	7.85	11.85
Social-VAE	(Xu, Hayet, and Karamouzas 2022)	8.10	11.72
SIT	(Shi et al. 2022)	8.59	15.27
MSRL	(Wu et al. 2023)	8.22	13.39
S-CSR*	(Zhou et al. 2021)	2.77	3.45
NSP-SFM*	(Yue, Manocha, and Wang 2022)	<u>1.78</u>	<u>3.44</u>
CSR*	(Zhou et al. 2023)	4.87	6.32
Ours*	-	**1.48**	**1.06**

Ablating the interaction-conditioned CVAE. In this experiment (named *Ours/wo*), we connect the coarse prediction model in SocialCVAE with the same CVAE model as (Zhou et al. 2021; Yue, Manocha, and Wang 2022; Zhou et al. 2023), which is only conditioned on the past trajectories, to learn the random residuals for the predicted preferred position from the coarse prediction model. Tab. 3 shows the quantitative results on SDD. Because the model doesn't consider pedestrian interactions and learns unexplainable motion randomness, compared with our full model, significant performance degradation occurs on both ADE and FDE, demonstrating the importance of our interaction-conditioned CVAE model for achieving better performance.

Ablating the attention model. In this experiment, we use the temporal extrapolation velocity generated by the temporal encoder as the output coarse preferred velocity of the coarse prediction model. The prediction results on SDD in Tab. 3 show performance degradation compared to our full model. However, when compared with the SOTA baselines, it still achieves better prediction accuracy, demonstrating the importance of our proposed interaction-conditioned CVAE model which learns the uncertainty of human motions.

Table 3: Ablation study of different components of our method on the SDD dataset. F_{goal} denotes the goal-attraction model proposed by NSP-SFM.

Components			ADE/FDE
F_{goal}	Attention model	Interaction -conditioned CVAE	
-	✓	✗	8.64/13.72
-	✗	✓	1.76/1.57
✓	✗	✓	1.56/2.71
-	✓	✓	**1.48/1.06**

Ablating the coarse prediction model. We also conduct another ablation experiment, named *GSocialCVAE*, by replacing the coarse motion prediction model in Sec. 3.3 with the goal-attraction model from the SOTA NSP-SFM method (Yue, Manocha, and Wang 2022), to further demonstrate the importance of the interaction-conditioned multimodal learning scheme employed in SocialCVAE. Tab. 3 gives the quantitative results of GSocialCVAE on SDD, showing performance degradation compared with our full model. However, when compared with NSP-SFM, GSocialCVAE achieves better performance with 11.80% improvement on ADE and 20.35% improvement on FDE, demonstrating the better prediction capability of our energy interaction-conditioned CVAE model for human motion uncertainty learning.

4.3 Qualitative Evaluation

We first visualize the predicted trajectories in several scenarios to illustrate the effectiveness of our method. The visualization results are shown in Fig. 5.

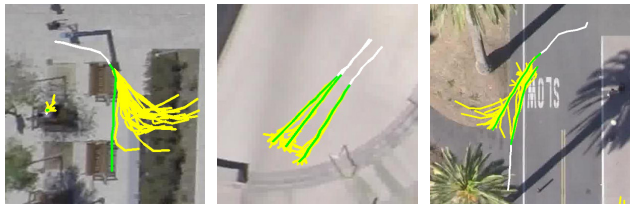
Predicted trajectory comparison. To further validate the better performance of our model, in Fig. 6, we compare our visualization results with the SOTA NSP-SFM model (Yue, Manocha, and Wang 2022). NSP-SFM may predict trajectories that obviously deviate from the ground-truth final positions. This is because NSP-SFM learns force-driven steering behaviors plus with unexplainable motion randomness;

(a) NSP-SFM-1 (b) NSP-SFM-2 (c) NSP-SFM-3

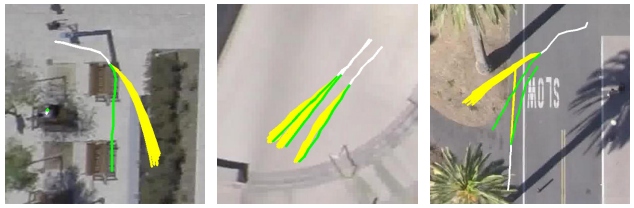
(d) GSocialCVAE-1 (e) GSocialCVAE-2 (f) GSocialCVAE-3

(g) Ours-1 (h) Ours-2 (i) Ours-3

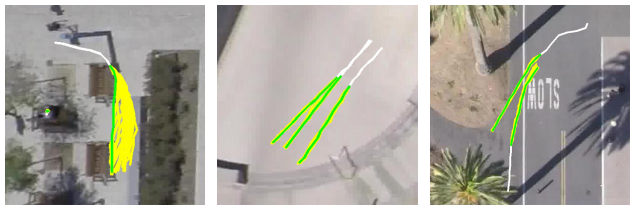
Figure 6: Visualization comparisons with NSP-SFM and GSocialCVAE. The visualized trajectories are the best predictions sampled from 20 trials. Our method predicts future trajectories closer to the ground truth than compared methods. The purple dots in the visualization results of NSP-SFM and GSocialCVAE represent the sampled goals.



(a) NSP-SFM-1 (b) NSP-SFM-2 (c) NSP-SFM-3



(d) Ours/wo-1 (e) Ours/wo-2 (f) Ours/wo-3



(g) Ours-1 (h) Ours-2 (i) Ours-3

Figure 7: Visualization comparisons of the multiple predicted trajectories with NSP-SFM and SocialCVAE without the interaction-conditioned CVAE (Ours/wo). Our method predicts more socially reasonable future trajectories than the compared methods. The white and green lines are the observed and ground-truth trajectories. The yellow lines for each pedestrian are the 20 predicted trajectories.

the predicted results show strong determinism in reaching a sampled final goal. In contrast, SocialCVAE employs an energy interaction-conditioned CVAE model for learning socially reasonable human motion uncertainty, thus achieving better prediction performance.

In Fig. 6, we also compare with the visualization results of the aforementioned ablation model GSocialCVAE. Due to the determinism nature of the goal-attraction model (Yue, Manocha, and Wang 2022), compared with our full model's result (Figs. 6g-6i), the predicted trajectories of GSocialCVAE show slight deviation from the ground-truth trajectories because the predicted goal is far from the ground truth. However, GSocialCVAE shows better visual results than NSP-SFM, demonstrating the proposed method's prediction capability to achieve better performance.

Interaction-conditioned multimodal prediction. As shown in Fig. 7, we compare the multiple predicted trajectories of the NSP-SFM, the ablation experiment of SocialCVAE without the interaction-conditioned CVAE (Ours/wo), and our full model (Ours). Our full model's results in Figs. 7g-7i demonstrate that by conditioning on the socially explainable interaction energy map, SocialCVAE learns better human motion uncertainty than the model without conditioned on interaction. Figs. 7h and 7i also demonstrate that SocialCVAE can predict socially reasonable trajectories for avoiding potential collisions than the models without conditioned on interaction.

5 Conclusion

In this work, we present SocialCVAE, a novel multimodal pedestrian trajectory prediction method with an interaction-conditioned CVAE model for learning socially reasonable human motion randomness. SocialCVAE explicitly models the anticipated social relationships of pedestrians and their neighbors by using an interaction energy map generated based on an energy-based interaction model. Taking the interaction energy map as a condition, the CVAE model can learn the uncertainty of human motions while maintaining social awareness. The proposed method outperforms existing state-of-the-art methods in achieving higher prediction accuracy. One limitation is that our method is computationally inefficient as we sequentially predict the energy map for each pedestrian. In the future, we will improve the computation performance by exploring other formulations of energy-based interaction.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 961–971.

Bae, I.; and Jeon, H.-G. 2021. Disentangled multi-relational graph convolutional network for pedestrian trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, 911–919.

Bisagno, N.; Zhang, B.; and Conci, N. 2018. Group lstm: Group trajectory prediction in crowded scenarios. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 213–225.

Duan, J.; Wang, L.; Long, C.; Zhou, S.; Zheng, F.; Shi, L.; and Hua, G. 2022. Complementary attention gated network for pedestrian trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 542–550.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2255–2264.

Guy, S. J.; Chhugani, J.; Curtis, S.; Dubey, P.; Lin, M. C.; and Manocha, D. 2010. PLEdestrians: A least-effort approach to crowd simulation. In *Symposium on Computer Animation*, 119–128.

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

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*, 9(8): 1735–1780.

Huang, Y.; Bi, H.; Li, Z.; Mao, T.; and Wang, Z. 2019. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 6272–6281.

Karamouzas, I.; Skinner, B.; and Guy, S. J. 2014. Universal power law governing pedestrian interactions. *Physical Review Letters*, 113(23): 238701.

Karamouzas, I.; Sohre, N.; Narain, R.; and Guy, S. J. 2017. Implicit crowds: Optimization integrator for robust crowd simulation. *ACM Transactions on Graphics (TOG)*, 36(4): 1–13.

Kothari, P.; Sifringer, B.; and Alahi, A. 2021. Interpretable social anchors for human trajectory forecasting in crowds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 15556–15566.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. In *Computer Graphics Forum*, volume 26, 655–664.

Mangalam, K.; An, Y.; Girase, H.; and Malik, J. 2021. From goals, waypoints & paths to long term human trajectory forecasting. In *International Conference on Computer Vision (ICCV)*, 15233–15242.

Mangalam, K.; Girase, H.; Agarwal, S.; Lee, K.-H.; Adeli, E.; Malik, J.; and Gaidon, A. 2020. It is not the journey but the destination: endpoint conditioned trajectory prediction. In *European Conference on Computer Vision (ECCV)*, 759–776.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 14424–14432.

Pellegrini, S.; Ess, A.; Schindler, K.; and Van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *International Conference on Computer Vision (ICCV)*, 261–268.

Ren, J.; Xiang, W.; Xiao, Y.; Yang, R.; Manocha, D.; and Jin, X. 2019. Heter-Sim: Heterogeneous multi-agent systems simulation by interactive data-driven optimization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 27(3): 1953–1966.

Reynolds, C. W. 1987. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, 25–34.

Reynolds, C. W.; et al. 1999. Steering behaviors for autonomous characters. In *Game Developers Conference*, volume 1999, 763–782.

Robicquet, A.; Sadeghian, A.; Alahi, A.; and Savarese, S. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European Conference on Computer Vision (ECCV)*, 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 1349–1358.

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

Shi, L.; Wang, L.; Long, C.; Zhou, S.; Zheng, F.; Zheng, N.; and Hua, G. 2022. Social interpretable tree for pedestrian trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 2235–2243.

Shi, L.; Wang, L.; Long, C.; Zhou, S.; Zhou, M.; Niu, Z.; and Hua, G. 2021. SGCN: Sparse graph convolution network for pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8994–9003.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems (Neurips)*, 30.

Vemula, A.; Muelling, K.; and Oh, J. 2018. Social attention: Modeling attention in human crowds. In *2018 IEEE Inter-*

national Conference on Robotics and Automation (ICRA), 4601–4607.

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

Xiang, W.; Wang, H.; Zhang, Y.; Yip, M. K.; and Jin, X. 2023. Model-based crowd behaviours in human-solution space. In *Computer Graphics Forum*, e14919.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 6498–6507.

Xu, P.; Hayet, J.-B.; and Karamouzas, I. 2022. Socialvae: Human trajectory prediction using timewise latents. In *European Conference on Computer Vision (ECCV)*, 511–528.

Yu, C.; Ma, X.; Ren, J.; Zhao, H.; and Yi, S. 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European Conference on Computer Vision (ECCV)*, 507–523.

Yuan, Y.; Weng, X.; Ou, Y.; and Kitani, K. M. 2021. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 9813–9823.

Yue, J.; Manocha, D.; and Wang, H. 2022. Human trajectory prediction via neural social physics. In *European Conference on Computer Vision (ECCV)*, 376–394.

Yue, J.; Monocha, D.; and Wang, H. 2023. Human trajectory forecasting with explainable behavioral uncertainty. arXiv:2307.01817.

Zhou, H.; Ren, D.; Yang, X.; Fan, M.; and Huang, H. 2021. Sliding sequential CVAE with time variant socially-aware rethinking for trajectory prediction. *arXiv preprint arXiv:2110.15016*.

Zhou, H.; Ren, D.; Yang, X.; Fan, M.; and Huang, H. 2023. CSR: cascade conditional variational auto encoder with socially-aware regression for pedestrian trajectory prediction. *Pattern Recognition*, 133: 109030.