Supplementary Material

January 23, 2025

1 Introduction

This supplementary material contains additional details that support the main paper. It includes proofs, data, derivations, and other information that are essential for understanding the work but are too detailed to include in the main paper.

2 Appendix A: Detailed Explanation of The Formulas

Here we present the detailed explanation on section 3.1 and section 3.2 in our main text.

2.1 Environment Constrained Social Interaction

To obtain high-dimensional environment constrained social interaction, we employ Social Representation Descriptor using an angle-based principle. Specifically, for the target pedestrian, its surrounding pedestrians and obstacles collectively form the social representations. Social Representation Descriptor comprises two components: Neighbor Pedestrians Perceptor and Surrounding Obstacle Perceptor.

Neighbor Pedestrian Perceptor. Given the observed trajectories of N pedestrians in a scene, Neighbor Pedestrian Perceptor can capture the interactions among the target pedestrian and his/her neighbors through the predefined rules. Existing methods use an angle-based principle to model social interactions. However, the method only focuses on the target pedestrian and overlooks the social interaction representations from the corresponding neighbor pedestrians. In this work, we take all pedestrians in a scene into account to accurately capture the social interactions. To comprehensively characterize each pedestrian's social interaction from various directions, we compute the angles of pedestrian j relative to the target pedestrian i by the following formula:

$$\theta_{ij} = \arctan\left(\frac{y_j^{T_h} - y_i^{T_h}}{x_j^{T_h} - x_i^{T_h}}\right), \tag{1}$$

where $arctan(\cdot)$ is the inverse tangent function that computes the angle of the auxiliary coordinate relative to the target. Here, θ_{ij} represents the relative motion direction of pedestrian j with respect to the target pedestrian i.

The characteristics of each directional area are defined by a combination of three human-defined rules: velocity, direction, and distance [15]. The velocity representations f^i_{vlc} of the target pedestrian i is calculated by using the Euclidean distance from starting position to ending of their observed trajectory. The direction representation f^i_{drt} of the target pedestrian i is expressed by the mean relative motion angles of all neighbor pedestrians. The distance representation f^i_{dst} is computed by the average distance from last position within the observed trajectory for each pedestrian to that of the target pedestrian. Formally,

$$\begin{aligned} f^i_{vlc} &= \frac{1}{N}\sum_{j=1}^{N}\left\|(x_j^{T_h}, y_j^{T_h}) - (x_j^1, y_j^1)\right\|_2 \quad (j \neq i), \\ f^i_{drt} &= \frac{1}{N}\sum_{j=1}^{N}\theta_{ij} \quad (j \neq i), \\ f^i_{dst} &= \frac{1}{N}\sum_{j=1}^{N}\left\|(x_j^{T_h}, y_j^{T_h}) - (x_i^{T_h}, y_i^{T_h})\right\|_2 \quad (j \neq i). \end{aligned} \tag{2}$$

Since a pedestrian can not simultaneously focus on every surrounding pedestrian in all directions, we divide the vicinity of the target pedestrian into p angular partitions, and calculate the social interaction representations $f^{p,i}_{nei}$ relative to the target pedestrian i in the direction of p. It is important to note that if no pedestrians are present within a specific directional area, the corresponding factors will be set to zero. Each partition's representation $f^{p,i}_{nei}$ is computed using the following formula:

$$f^{p,i}_{nei} = \text{Concat}(f^{p,i}_{vlc}, f^{p,i}_{drt}, f^{p,i}_{dst}). \tag{3}$$

The final social interaction features f^i_{si} of pedestrian i is a combination of different partition's representations which can be represented as:

$$f^i_{si} = \text{Concat}(f^{1,i}_{nei}, f^{2,i}_{nei}, \cdots, f^{p,i}_{nei}). \tag{4}$$

Surrounding Obstacle Perceptor. Similar to Neighbor Pedestrian Perceptor, Surrounding Obstacle Perceptor also follows angle-based principle to extract surrounding obstacle representations. First, we convert the scene segmentation map into pixel-level occupancy grid map with a shape of $(100, 100)$. Next, a homography matrix provided by the dataset is applied to transform the target's 2D coordinate trajectories into pixel trajectories. Pixels with a value of 0 are treated as obstacles, and an angle-based approach is used to compute the three components of surrounding obstacle features: velocity, direction, and distance. The resulting surrounding obstacle features are denoted as f^i_{so}. Formally:

$$f^i_{so} = \text{Concat}(0, f^i_{\hat{drt}}, f^i_{\hat{dst}}), \tag{5}$$

here $f^i_{\hat{drt}}$ and $f^i_{\hat{dst}}$ are the surrounding direction and distance representations of the target pedestrian i. Ultimately, we obtain each pedestrian's social interactions f^i_{si} and surrounding obstacles f^i_{so} by jointly composing social representations f^i_{sr}, which is computed by the following formula:

$$f^i_{sr} = \text{MLP}(\tanh(\text{Concat}(f^i_{si}, f^i_{so}))). \tag{6}$$

2.2 Motion Pattern Descriptor

The Motion Pattern Descriptor leverages SVD to identify and retain essential low-dimensional features from trajectory data. Through decomposing the observed and predicted matrices into their singular components, we can effectively capture the underlying motion patterns of pedestrians.

SVD is a fundamental matrix decomposition measure with a wide range of applications in fields, such as recommendation systems, image compression, and signal processing [4]. Given a matrix A, SVD decomposes it into three matrices by the following formula:

$$A \quad = \quad U\Sigma V^T, \tag{7}$$

where U is an orthogonal matrix consisting of the eigenvectors of A^TA, and the column vectors of U are referred to the left singular vectors. Similarly, V is an orthogonal matrix made up of the eigenvectors of A^TA. The column vectors of V are known as the right singular vectors. The matrix Σ is a diagonal matrix with non-zero values on its diagonal which are the singular values arranged in descending order. The number of non-zero singular values corresponds to the rank of the matrix A, which is not greater than the minimum of the number of rows and columns in A.

To effectively extract representative motion patterns of pedestrians, we apply SVD to decompose the observed matrix H and the predicted matrix F from the training dataset. The decomposition process is expressed as follows:

$$\begin{aligned} H &= U_h \Sigma_h V_h^T, \\ F &= U_f \Sigma_f V_f^T, \end{aligned} \tag{8}$$

where $U = [U_1, U_2, \cdots, U_{2h}]$ and $V = [V_1, V_2, \cdots, V_N]$, and Σ is a diagonal matrix containing m singular values. Here, h and f represent the length of the observed and predicted trajectories, respectively. The singular values in Σ are all greater than 0, indicating the relative weight of each motion pattern.

To select the most representative motion patterns and enhance the model's robustness, we choose the top r singular eigenvectors as the pedestrian's representative motion patterns, and further improve the model's generalization ability. We then use a linear combination of these r left singular vectors to approximate the trajectory. Formally,

$$\begin{aligned} \tilde{H}_n &= U_{h,r} C_{h,n}, \\ \tilde{F}_n &= U_{f,r} C_{f,n}, \end{aligned} \tag{9}$$

where $U_{h,r}$ represents the past motion patterns comprising r selected eigenvectors, and $C_{h,n}$ denotes the correlation coefficients of the past trajectories, which determine the velocity and direction of the past motion patterns. The same approach is applied to approximate the future trajectory $\tilde{F}_n$.

3 Appendix B: Additional Experiments

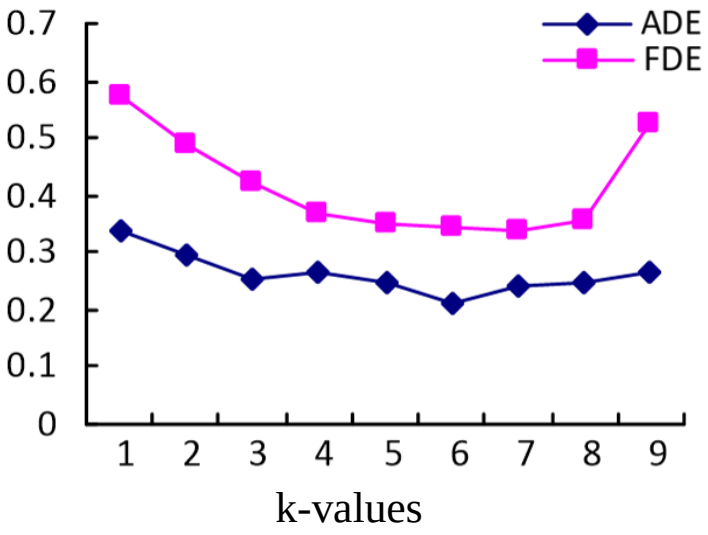

Figure 1: Visualization of effectiveness when setting different values of iteration k on ETH-UCY dataset with SMP-SGCN.

The number of iterations k. To determine the appropriate value of k in additive fusion mechanism, we conduct an ablation study to assess the effect of the hyperparameter $k(k \in [1, 9])$ on the ETH-UCY dataset with SMP-SGCN model. As shown in Fig. 1, as the value of k increases, both ADE and FDE initially decrease and then increase. Specifically, ADE reaches its minimum when $k = 6$, while FDE reaches its minimum when $k = 7$. Moreover, FDE tends to stabilize while $k = 5$, 6, and 7. To simultaneously achieve the best performance on ADE and FDE as much as possible, we select $k = 6$ as the most appropriate value. This result indicates that the additive fusion mechanism not only enhances the social awareness of motion patterns but also avoid overfitting problem when setting k as 6.

Table 1: Comparison of Computational Overhead and Prediction Accuracy

	SGCN vs SMP-SGCN	**STGCNN vs SMP-STGCNN**	**Implicit vs SMP-Implicit**
Training convergence time	1h31min vs 1h49min	1h36min vs 1h54min	1h53min vs 2h14min
ADE/FDE↓	0.35/0.63 vs **0.21/0.34**	0.45/0.75 vs **0.22/0.37**	0.33/0.67 vs **0.21/0.36**

Computational Cost: As shown in Table 1 (ADE/FDE values extracted from Table 1 in our manuscript), after integrating the SocialMP module into baseline models, we can see that although the training cost indeed increases slightly, the corresponding prediction accuracy has been improved significantly.

4 Appendix C: Baseline Models

Baseline models. We integrate our SocialMP into the following baseline models: *SGCN* [14], *STGCNN* [10] and *Implicit* [11] to validate the effectiveness of our proposed method. Due to different state-of-the-art trajectory prediction methods using various datasets, for fair comparison, we selected different baselines for ETH-UCY and SDD datasets. Specifically, we compare SocialMP models with *Social-LSTM* [1], *Social-GAN* [5], *PECNet* [8], *Trajectron++* [13], *STGAT* [6], *AgentFormer* [18],*GroupNet* [17], *GP-Graph* [3], *Graph-TERN* [2] and *SMEMO* [9] on ETH-UCY dataset. On SDD, we compare our model on the metrics of ADE and FDE with *SocialGAN* [5], *Sophie* [12], *PECNet* [8], *BCDiff* [7], *Graph-TERN* [2], *MRL* [16] and *SMEMO* [9].

5 Appendix D: Implementation Details

Implementation Details. All experiments were conducted on an Nvidia Tesla V100 GPU with 32GB of memory. The training process utilized the Adam optimizer with a learning rate of 0.001. The number of epochs was set to 256, and a batch size of 128 was employed for efficient processing. To prevent overfitting, a weight decay of 0.0001 was applied.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, Las Vegas, 2016. IEEE Computer Society.

[2] Inhwan Bae and Hae-Gon Jeon. A set of control points conditioned pedestrian trajectory prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(690):6155 – 6165, 2023.

[3] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Learning pedestrian group representations for multi-modal trajectory prediction. In *Proceedings of the European Conference on Computer Vision*, volume 13682, page 270–289, Tel-Aviv, Israel, 2022. Springer, Cham.

[4] Inhwan Bae, Jean Oh, and Hae-Gon Jeon. Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 9983–9995, Los Alamitos, 2023. IEEE Computer Society.

[5] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2255–2264, Los Alamitos, CA, USA, 2018. IEEE Computer Society.

[6] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6271–6280, Berlin, Heidelberg, 2019. Springer-Verlag.

[7] Rongqing Li, Changsheng Li, Dongchun Ren, Guangyi Chen, Ye Yuan, and Guoren Wang. Bcdiff: Bidirectional consistent diffusion for instantaneous trajectory prediction. In *Advances in Neural Information Processing Systems*, volume 36, pages 14400–14413, Louisiana,United States, 2023. Curran Associates, Inc.

[8] 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 *Computer Vision – ECCV 2020*, pages 759–776, Cham, 2020. Springer International Publishing.

[9] F. Marchetti, F. Becattini, L. Seidenari, and A. Bimbo. Smemo: Social memory for trajectory forecasting. *IEEE Transactions on Pattern Analysis ; Machine Intelligence*, 46(06):4410–4425, 2024.

[10] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14412–14420, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[11] Abduallah Mohamed, Deyao Zhu, Warren Vu, Mohamed Elhoseiny, and Christian Claudel. Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, page 463–479, Tel Aviv, Israel, 2022. Springer-Verlag.

[12] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1349–1358, Long Beach,CA, USA, 2019. Piscataway, NJ: IEEE.

[13] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII*, page 683–700, Berlin, Heidelberg, 2020. Springer-Verlag.

[14] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. Sgcn:sparse graph convolution network for pedestrian trajectory prediction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8990–8999, Nashville, TN, USA, 2021. IEEE.

[15] Conghao Wong, Beihao Xia, Ziqian Zou, Yulong Wang, and Xinge You. Socialcircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19005–19015, Seattle, WA, USA, 2024. IEEE.

[16] Yuxuan Wu, Le Wang, Sanping Zhou, Jinghai Duan, Gang Hua, and Wei Tang. Multi-stream representation learning for pedestrian trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2875–2882, Washington, DC, USA, 2023. Curran Associates, Inc.

[17] C. Xu, M. Li, Z. Ni, Y. Zhang, and S. Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6488–6497, Los Alamitos, CA, USA, 2022. IEEE Computer Society.

[18] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2021. IEEE.