Who Walks With You Matters: Perceiving Social Interactions with Groups for Pedestrian Trajectory Prediction

Supplementary Material

A. Additional Quantitative Analysis

We only report the GPCC model's performance with part of the quantitative results due to page limitations. As the qualitative analysis shown in the main manuscript, the proposed GPCC model presents the capability to handle different prediction scenes. This section further validates the model's effectiveness by presenting additional quantitative analysis of the datasets qualitatively analyzed in the manuscript.

A.1. Experimental Configurations

NBA comprises trajectories of both players and basketball captured by SportVU tracking systems during NBA games. Following the methodology proposed by Xu et al. [13, 14], we set the parameters to be $\{n_p, n_f, T\} = \{5, 10, 0.4\text{s}\}$, randomly selecting approximately 50000 samples (ego trajectories), with 65% allocated for training, 25% for testing, and 10% for validation.

nuScenes contains 1000 driving scenes collected in the urban area of Boston and Singapore. Each scene is 20 seconds long and is annotated at a rate of 2 frames per second. In the manuscript, we only use the two-dimensional trajectories of vehicles to evaluate our GPCC model. We follow the methodology proposed by [3] of $\{n_p = 4, n_f = 12, T = 0.5\text{s}\}$, and training strategy proposed by [7] of using 550 scenes to train, 150 scenes to validate, and the other 150 scenes to test.

Model	$n_p = 5$	$n_f = 10$
Social-LSTM[1] (2016)	0.88/1.53	1.79/3.16
S-GAN[2] (2018)	0.85/1.36	1.62/2.51
Social-STGCNN[6] (2020)	0.75/0.99	1.59/2.37
GroupNet+NMMP[13] (2022)	0.69/1.08	1.25/1.80
GroupNet+CVAE[13] (2022)	**0.62**/0.95	**1.13**/1.69
SocialCircle[10] (2024)	0.67/0.90	1.18/**1.46**
GPCC **(Ours)**	**0.62/0.87**	1.19/1.58

Table S1. Comparisons on NBA under $\{n_p, n_f, T\} = \{5, 10, 0.4\text{s}\}$. Metrics are "ADE/FDE" in meters under *best-of-20* on $5, 10$ future steps, and lower ADE and FDE indicate better performance.

A.2. Analysis

NBA players on the NBA dataset interact with each other differently compared with ETH-UCY and SDD. In Tab. S1, we can observe that the GPCC model outperforms other state-of-the-art methods when $n_p = 5$ and reaches the second best in FDE when $n_p = 10$. The results of NBA validate the GPCC model's capability of modeling different social interactions.

Model	*best-of-5*	*best-of-10*
Trajectron++[8] (2020)	3.14/7.45	2.46/5.65
Y-net[5] (2020)	2.46/5.15	1.88/3.47
AF[15] (2021)	1.59/3.14	1.30/2.47
E-V^2-Net[12] (2023)	1.46/3.18	1.15/2.37
SocialCircle[10] (2024)	1.44/3.10	1.13/2.30
MUSE-VAE[3] (2022)	1.38/**2.90**	1.09/**2.10**
GPCC **(Ours)**	**1.33**/2.94	**1.08**/2.27

Table S2. Comparisons on nuScenes under $\{n_p, n_f, T\} = \{4, 12, 0.5\text{s}\}$. Metrics are "ADE/FDE" in meters under *best-of-k* ($k = 5, 10$) and lower values indicate better prediction performance.

We only consider trajectories of vehicles only on nuScenes. Interactions between vehicles could differ entirely from those between pedestrians, and there might not be such *group* relations between them. This property of nuScenes are discussed in Sec. 4.5. However, the GPCC model still gained a considerable prediction performance, as shown in Tab. S2.

Although results in Tab. S4 are relatively minor compared to those in Tab. 1, we could still see a performance drop of larger ADE and FDE at v1,v2 and v3. We could observe that agents in SDD tend to move and behave independently, and there seems to be less interaction between agents from the relatively high eye-bird view compared to scenes in ETH-UCY, and this might be the reason for the smaller contributions of the Group method and Conception module of the proposed GPCC model when evaluating on the SDD dataset.

In Tab. S3, we could observe that the performance drops the most when disabling both the Group method and Conception module on the NBA dataset. However, the prediction accuracy when disabling the Conception module only (v1) reaches the same level as the original GPCC model (discussed in Sec. 4.5). The results demonstrate an opposite performance change on the nuScenes dataset. The prediction performance drops when disabling the Conception module (v1) or both of them (v3) and keeps the same level as the original GPCC model when disabling the Group method only (v2). The reason for this phenomenon might be common sense that vehicles moving on the road are not in distinct groups. It might lead the model to learn information through group relations between cars if we use the

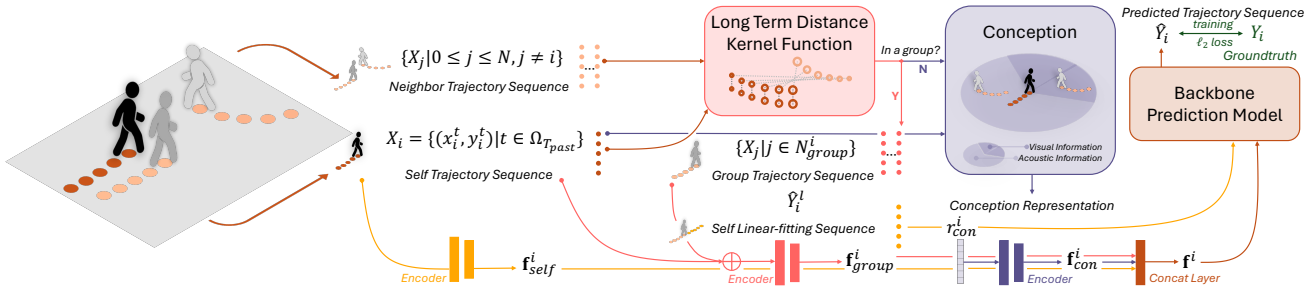

Figure S1. More detailed version of the schema of the modeling process of the proposed GPCC model and the computation pipeline of the Group method and the Conception module.

ID	Group	Conception	NBA		ΔNBA		nuScenes		ΔnuScenes	
			$n_p = 5$	$n_f = 10$	$n_p = 5$	$n_f = 10$	*best-of-5*	*best-of-10*	*best-of-5*	*best-of-10*
v1	•	○	0.62/0.87	1.19/1.58	0.0%/0.0%	0.0%/0.0%	1.38/3.05	1.10/2.30	3.8%/3.7%	1.9%/0.9%
v2	○	•	0.66/0.94	1.25/1.68	6.5%/8.0%	5.0%/6.3%	1.34/2.95	1.07/2.27	0.8%/0.3%	-0.9%/0.0%
v3	○	○	0.70/1.03	1.34/1.86	12.9%/18.4%	12.6%/17.7%	1.34/2.95	1.09/2.28	0.8%/0.3%	0.9%/0.4%
v0	•	•	0.62/0.87	1.19/1.58	0.0%/0.0%	0.0%/0.0%	1.33/2.94	1.08/2.27	0.0%/0.0%	0.0%/0.0%

Table S3. Ablation studies on NBA and nuScenes. "•" means using the corresponding method or module and "○" indicates the opposite. "ID" represents a different variation index of the proposed GPCC model, and "ΔNBA, nuScenes" indicates the percentage of **performance drops** compared to the full GPCC model(v0).

ID	Group	Conception	SDD	ΔSDD
v1	•	○	6.44/10.34	0.8%/1.7%
v2	○	•	6.46/10.38	1.1%/2.1%
v3	○	○	6.40/10.17	0.2%/0%
v0	•	•	6.39/10.17	0.0%/0.0%

Table S4. Ablation studies on SDD. "•" means using the corresponding method or module and "○" indicates the opposite. "ID" represents a different variation index of the proposed GPCC model, and "ΔSDD" indicates the percentage of **performance drops** compared to the full GPCC model(v0).

Group method on nuScenes.

B. Additional Analysis of Time Efficiency

Pedestrian trajectory prediction task requires low-latency prediction performance to be integrated into corresponding applications, *e.g.*, autonomous driving. We also use Apple Mac Studio (M2 Max) to evaluate the time efficiency of different methods, as shown in Sec. 4.4. Considering the practical amount of pedestrians moving in a multi-agent scene, we evaluate the average inference time of 100 target agents (batchsize 100) using different methods.

The proposed GPCC model demonstrates considerable inference speed compared with other methods using Transformer as part of backbone prediction model [10] (32ms with batchsize 100), [11] (96ms with batchsize 100). With the inference time already satisfying handling 99 more agents between two adjacent intervals, the model should meet the low-latency requirement of the trajectory prediction task [4].

C. Additional Analysis of FOV Partitions

As represented in Tabs. S5 and S6, we first conduct variation experiments on pedestrian dataset ETH-UCY and game dataset NBA to observe how the performance of the proposed GPCC model vary with FOV angle θ_{FOV}.

In Tab. S5, the best performance comes at the original model(v0) whose θ_{FOV} is set to be $180°$. This result is in line with the previous study [9] that a human's single-eye FOV angle is around $150°$ and the combined FOV from both eyes reaches about $200°$, which is why we chose to set the $\theta_{\text{FOV}} = 180°$ at the first place. Although we can observe that the performance of the GPCC model drops little from the overall results on the whole ETH-UCY dataset, the prediction performance on specific subsets such as univ, zara1, and zara2 drops relatively more than the other subsets. This might be aroused that there are more social interactions in these subsets and the change of θ_{FOV} modifies how the Conception module perceives social interactions with other agents.

Things are getting more interesting in the NBA dataset shown in Tab. S6. It can be observed that the prediction performance is becoming better from $\theta_{\text{FOV}} = 90°$ to $\theta_{\text{FOV}} = 360°$ (v5, v6, v7 and v8). NBA players might need to spread their attention to a broader FOV to gain more information on the court and make decisions of move-

ID	θ_{FOV}	eth	hotel	univ	zara1	zara2	ETH-UCY	ΔETH-UCY
v4	0°	0.26/0.40	0.10/0.15	0.26/0.47	0.18/0.30	0.13/0.25	0.19/0.31	5.6%/6.9%
v5	90°	0.25/0.38	0.11/0.16	0.26/0.45	0.18/0.30	0.14/0.23	0.19/0.30	5.6%/3.4%
v6	135°	0.26/0.39	0.10/0.16	0.26/0.46	0.18/0.30	0.14/0.23	0.19/0.31	5.6%/6.9%
v7	270°	0.26/0.41	0.11/0.17	0.26/0.46	0.18/0.31	0.14/0.23	0.19/0.32	5.6%/10.3%
v8	360°	0.26/0.40	0.10/0.15	0.27/0.49	0.18/0.31	0.14/0.23	0.19/0.32	5.6%/10.3%
v0	180°	0.25/0.38	0.10/0.15	0.25/0.44	0.17/0.28	0.13/0.22	0.18/0.29	0.0%/0.0%

Table S5. FOV angle θ_{FOV} analysis on EHT-UCY dataset. Variation "ID" is continuous from Tab. 3.

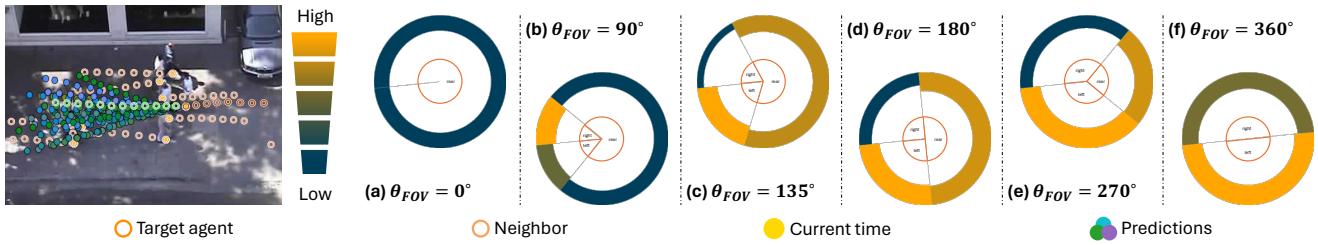

Figure S2. Visualization of attention value in the form of concentric fan chart varying with FOV angle θ_{FOV}. The model pays more attention to wider regions, whose color tends to be yellow accordingly.

ID	θ_{FOV}	NBA		ΔNBA	
		$n_p = 5$	$n_f = 10$	$n_p = 5$	$n_f = 10$
v4	0°	0.62/0.86	1.19/1.57	0.0%/-1.1%	-0.8%/-1.3%
v5	90°	0.64/0.89	1.22/1.62	3.2%/2.3%	2.5%/1.9%
v6	135°	0.63/0.88	1.20/1.60	1.6%/1.1%	0.8%/1.3%
v7	270°	0.62/0.86	1.19/1.58	0.0%/-1.1%	0.0%/0.0%
v8	360°	0.62/0.87	1.20/1.59	0.0%/0.0%	0.8%/0.6%
v0	180°	0.62/0.87	1.19/1.58	0.0%/0.0%	0.0%/0.0%

Table S6. FOV angle θ_{FOV} analysis on NBA dataset. Variation "ID" is continuous from Tab. S3.

ments based on this information. Furthermore, when the $\theta_{\mathrm{FOV}} = 0°$, the Conception module perceives interactions all the same by only considering the distance factor of other agents, which might lead to surprisingly minor improvements in the prediction performance.

We further visualize the attention value of the Conception module at different FOV angle settings with different concentric fan charts as shown in Fig. S2. By comparing charts when $\theta_{\mathrm{FOV}} = 90°$ and $\theta_{\mathrm{FOV}} = 135°$ (Fig. S2 (b) and (c)), we can observe a change of relative attention value in right and left partitions. When using a wider FOV angle, agents divided into rear partitions at a narrower FOV angle can be included into left or right partitions so that they can be paid more attention.

D. Additional Analysis of Choosing the Long-term Distance Threshold

In the main manuscript, we introduce the Group method and its core component, long-term distance kernel function $K(\cdot)$. When calculating the long-term distance kernel function, we mention the threshold to determine whether the agent belongs to the same group as the target agent without detailedly introducing how to choose the threshold d_{m} due to page limitations. This section demonstrates how we determine the kernel function's threshold d_{m}.

As shown in the manuscript, we also use the family group here as an example, considering Child as the target agent. We also use this example because this family example includes diverse scenes from multiple neighboring agents to no one else around. Members of the group (Mother, Father, and Child) behave differently according to their own will. Overall, Child seems to be talking with Father all the way from the Zara store to the other side of the road. Mother walks behind Child at a relatively larger distance compared to Father. In Fig. S3 (c1), we can observe that Father turns around to Mother to say something (at frame No.790), which further validates their grouping relations.

The long-term distance value of each agent is shown in the bar charts on the right. In Fig. S3 (c1),(d1),(e1), and (f1), the long-term distance of Father-Child or Mother-Child (marked in deep green) is distinctively lower than other agents (marked in deep blue). However, in Fig. S3 (a1), we can observe that the long-term distance of Mother ranks third among all neighboring agents of the target agent Child while Father is still at the top of the list. When plenty of agents exist in the scene, the long-term distance can be near each other when the time window used to calculate the distance sum is relatively short. Although an "unrelated" neighbor is classified as a group member, the trajectories seem reasonable in this condition. Further, we humans also

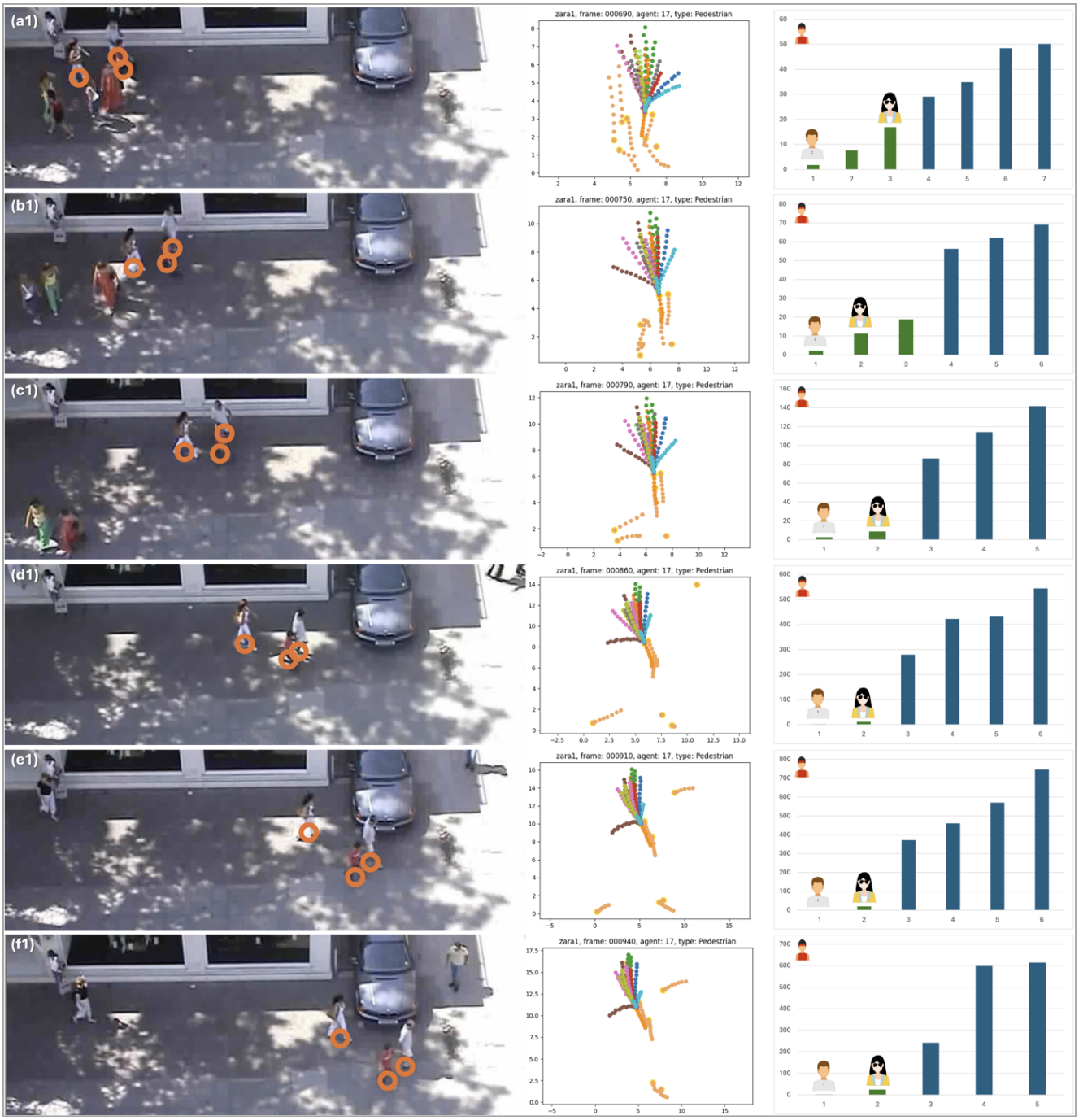

Figure S3. Illustration of how we design the threshold d_{m} of the long-term distance kernel function based on the sum of distance in bar charts corresponding to Fig. 3.

could not tell the groundtruth grouping relations by simply observing the coordinates information during T_{past} shown in the middle of Fig. S3.

Based on what we calculate among the ETH-UCY dataset, we design the threshold d_{m} to be 20. Despite occasional errors, it can exclude the near "unrelated" agents and the off-centered "related" agents at the same time.

E. Additional Analysis of Contribution Ratio

The visualization results in this manuscript present the contribution ratio of self, group, and contribution in a concen-

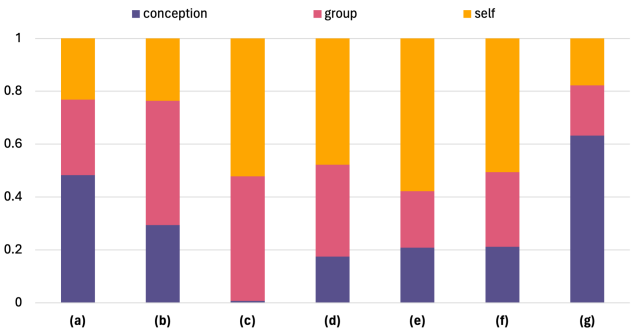

Figure S4. Contribution ratio of the **Conception** feature, the **Group** feature and the **Self** feature in the form of bar charts. Each bar corresponds to Fig. 5 in the main manuscript.

tric fan chart form. By comparing the angle of each concentric fan, we can observe which feature contributes the most and which plays little role in predicting future trajectories. Here, we further visualize this contribution ratio in a bar chart form, which can present a more accurate difference in contribution ratio. Fig. S4 (g) represents the largest contribution ratio in conception. This aligns with the situation that Fig. S4 (g) stands for the NBA dataset, where abundant interactions can be observed. Combining the fan charts in the manuscript and the bar charts here (Fig. S4), we can better understand how pedestrians' decisions to make movements originate from these three features.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 1

[2] 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 Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 1

[3] Mihee Lee, Samuel S Sohn, Seonghyeon Moon, Sejong Yoon, Mubbasir Kapadia, and Vladimir Pavlovic. Muse-vae: Multi-scale vae for environment-aware long term trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2221–2230, 2022. 1

[4] Shijie Li, Yanying Zhou, Jinhui Yi, and Juergen Gall. Spatial-temporal consistency network for low-latency trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1940–1949, 2021. 2

[5] 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. 1

[6] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020. 1

[7] Saeed Saadatnejad, Yi Zhou Ju, and Alexandre Alahi. Pedestrian 3d bounding box prediction. *arXiv preprint arXiv:2206.14195*, 2022. 1

[8] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 683–700. Springer, 2020. 1

[9] Hermann Von Helmholtz. *Handbuch der physiologischen Optik*. Voss, 1867. 2

[10] Conghao Wong, Beihao Xia, Ziqian Zou, Yulong Wang, and Xinge You. Socialcircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19005–19015, 2024. 1, 2

[11] Conghao Wong, Beihao Xia, Ziqian Zou, and Xinge You. Socialcircle+: Learning the angle-based conditioned interaction representation for pedestrian trajectory prediction. *arXiv preprint arXiv:2409.14984*, 2024. 2

[12] Beihao Xia, Conghao Wong, Duanquan Xu, Qinmu Peng, and Xinge You. Another vertical view: A hierarchical network for heterogeneous trajectory prediction via spectrums. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[13] 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 (CVPR)*, pages 6498–6507, 2022. 1

[14] 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 (CVPR)*, pages 6488–6497, 2022. 1

[15] 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 (ICCV)*, pages 9813–9823, 2021. 1