Supplementary Material of IPCC-TP: Utilizing Incremental Pearson Correlation Coefficient for Joint Multi-Agent Trajectory Prediction

Dekai Zhu[1,*], Guangyao Zhai[1,*], Yan Di[1,†], Fabian Manhardt[2],
Hendrik Berkemeyer[3,4], Tuan Tran[3], Nassir Navab[1], Federico Tombari[1,2], and Benjamin Busam[1,5]

[1] Technical University of Munich [2] Google
[3] Robert Bosch GmbH [4] University of Osnabrueck [5] 3Dwe.ai

1. Metrics

In our experiments, we leverage the *Minimum Joint Average Displacement Error* (**minJointADE**) and *Minimum Joint Final Displacement Error* (**minJointFDE**) as the evaluation metrics, which are specifically proposed for multi-agent trajectory prediction task in INTERPRET Challenge [3]. They are different from *Minimum Average Displacement Error* (minADE) and *Minimum Final Displacement Error* (minFDE), which are frequently used in the benchmarking of ego-motion prediction. Since we focus on predicting a scene-compliant future interaction among the agents, metrics that consider all agents in the scene, such as minJointADE and minJointFDE, are more suitable options.

MinJointADE represents the minimum value of the Euclidean Distance averaged by time and all agents between the ground truth and the mode with the lowest value [1]. Note that in the problem statement in subsection 3.1, we only analyze the case when the number of modes $M = 1$. In this case, the future states and the corresponding estimations at step t are $\mathcal{F}_t = \{\mathcal{F}_i^t | i = 1, ..., N\}$ and $\hat{\mathcal{F}}_t = \{\mathcal{F}_i^{\hat{t}} | i = 1, ..., N\}$ respectively, where $\mathcal{F}_i^t = \{\mathcal{F}_i^{tx}, \mathcal{F}_i^{ty}\}$ and $\mathcal{F}_i^{\hat{t}} = \{\mathcal{F}_i^{\hat{tx}}, \mathcal{F}_i^{\hat{ty}}\}$ are the position ground truth and estimation for agent i at this step. Since a specific current situation could develop into multiple possible future interactions, most MTP model predicts multiple interaction modes, where $M > 1$, thus $\mathcal{F}_i^{\hat{t}} = \{\mathcal{F}_{im}^t | m = 1, ..., M\}$ and $\mathcal{F}_{im}^{\hat{t}} = \{\mathcal{F}_{im}^{\hat{tx}}, \mathcal{F}_{im}^{\hat{ty}}\}$. The minJointADE is calculated as

$$\text{minJ-ADE} = \min_{1 \leq m \leq M} \frac{1}{NT} \sum_{i,t} \sqrt{(\mathcal{F}_{im}^{\hat{tx}} - \mathcal{F}_i^{tx})^2 + (\mathcal{F}_{im}^{\hat{ty}} - \mathcal{F}_i^{ty})^2}. \tag{1}$$

MinJointFDE represents the minimum value of the euclidean distance at the last predicted timestamps averaged by all agents between the ground truth and the mode with the lowest value [1]. The minJointFDE is defined as

$$\text{minJ-FDE} = \min_{1 \leq m \leq M} \frac{1}{N} \sum_{i} \sqrt{(\mathcal{F}_{im}^{\hat{Tx}} - \mathcal{F}_i^{Tx})^2 + (\mathcal{F}_{im}^{\hat{Ty}} - \mathcal{F}_i^{Ty})^2}. \tag{2}$$

2. Experiment Details

2.1. Data Preprocessing

NuScenes. In the nuScenes [2] dataset, the observable histories and predicted futures respectively last for 2s and 6s, with a sample rate of 2Hz. Thus the histories contain 4 time steps, and the futures contain 12 time steps. In the data preprocessing for Agentformer [6] and AutoBots [4], we remove agents with recorded past trajectories less than 2 steps or with incomplete future trajectories.

Argoverse 2. In the Argoverse 2 [5] dataset, the observable histories and predicted futures are 5s and 6s with a sample rate of 10Hz. Henceforth, the histories contain 50 time steps, and the futures contain 60 time steps. Compared with nuScenes, Argoverse 2 has significantly more agents in most scenes, and most agents have complete future trajectories. Thus, we only select agents with complete trajectories in the data preprocessing for Agentformer and AutoBots.

2.2. Training Details

For IPCC-TP and the backbones of Agentformer and AutoBots, we set a dropout rate of 0.1. We train the Agentformer for 50 epochs with an initial learning rate of 1e-4 in the evaluation on both nuScenes and Argoverse 2. We further decay the learning rate by 0.5 every 10 epochs. As for the training of AutoBots, we set the initial learning rate as 7.5e-4. In the evaluation on nuScenes, we train the model for 100 epochs. Thereby, in the first 50 epochs, we decay the learning rate by 0.5 every 10 epochs. In the evaluation on Argoverse 2, we train the model for 100 epochs, reducing the learning rate by 0.5x every 20 epochs. We train the models on a single NVIDIA Titan Xp.

[*] Equal contribution. [†] Corresponding author.

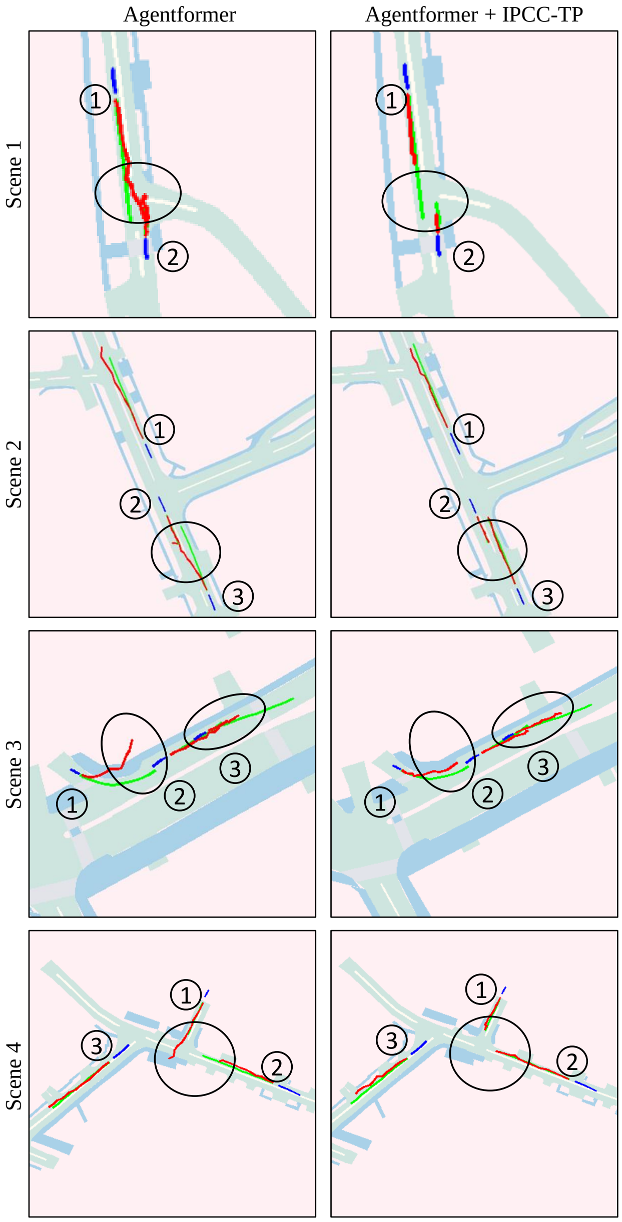

Figure 1. Comparisons between Agentformer and Agentformer + IPCC-TP on nuScenes (blue line: past trajectories, green line: ground truth, red line: predictions). **Scene 1**: Vehicle ① and ② drive in different directions, and they should keep their lanes without interfering with each other. **Scene 2**: Vehicle ①, ② and ③ should keep their lanes. **Scene 3**: Vehicle ① should turn left properly and then follow vehicle ② and ③. **Scene 4**: Vehicle ① should wait until vehicle ② passes the T-junction. Black circles show that the results of the Agentformer baseline are abnormal, while they become reasonable after the model is enhanced with IPCC-TP.

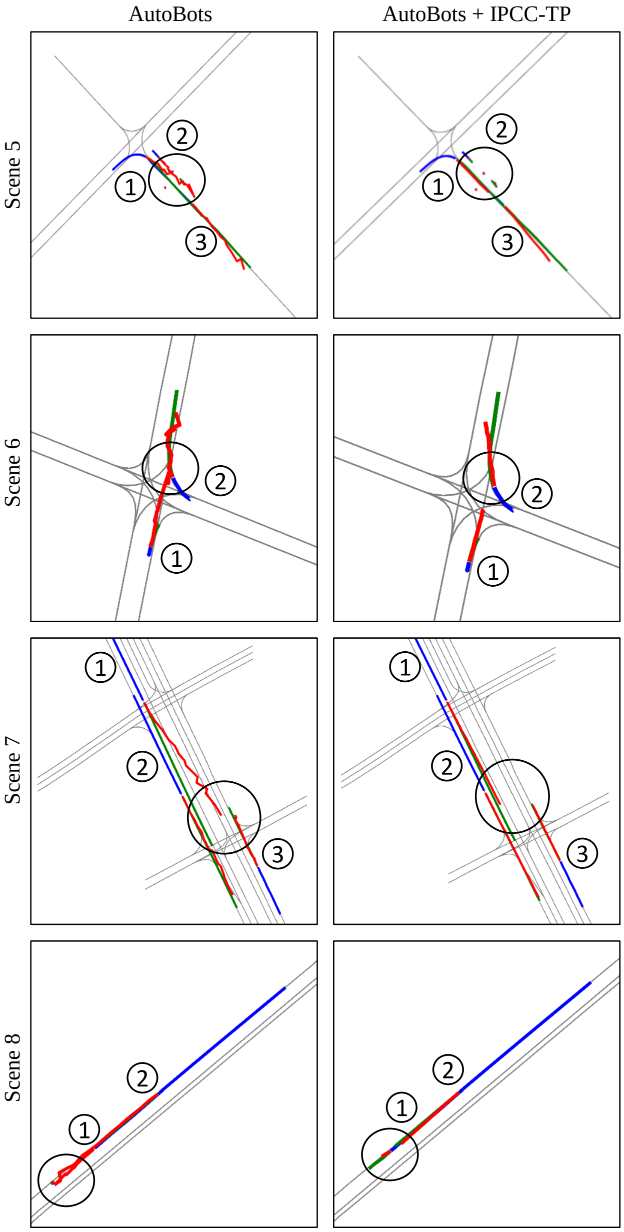

Figure 2. Comparisons between AutoBots and AutoBots + IPCC-TP on Argoverse 2 (blue line: past trajectories, green line: ground truth, red line: predictions, grey line: road center lines). **Scene 5**: Vehicle ① and ③ should keep their lanes while vehicle ② is about to stop on the side of the road. **Scene 6**: Vehicle ② is turning right and vehicle ① should yield to vehicle ②. **Scene 7**: Vehicle ①, ②, ③ should keep their lanes without interfering with each other. **Scene 8**: Vehicle ② is approaching vehicle ① while the latter starts moving. Black circles show that the results of AutoBots baseline are abnormal, while they become reasonable after the model is enhanced with IPCC-TP.

3. More Experiment Results

In this supplementary material, we first provide more results about the qualitative comparisons between the baseline

AutoBots+IPCC-TP	minJ-ADE(6)	minJ-FDE(6)
MLP Block	2.26	4.84
Attention Block	**2.20**	**4.79**

Table 1. Ablation study of Attention Block on Argoverse 2.

models and the enhanced models on nuScenes and Argoverse 2. Then, to support our statement about yaw angle θ estimation mentioned in Sec.3.3. IPCC Projection, we also study the error caused by the approximation in a quantitative way. Next, we provide the ablation study results on the Attention Block introduced in the main paper. Finally, we provide the result of AutoBots enhanced by our module compared with the original version on the INTERACTION dataset [7].

3.1. Qualitative Results on nuScenes and Argo 2

We show the comparison in Figure. 1 and Figure. 2. In addition, we visualize IPCC matrices in several scenes in nuScenes. The result is shown in Figure. 3.

3.2. Yaw Angle Analysis

As described in Sec.3.3 in the main paper, for short-term trajectory prediction (no longer than 6s), vehicles are unlikely to have sharp turns in such a short time, thus the angle θ_i^t based on the incremental movement is close to the actual yaw angle ϕ_i^t, and $\hat{\mathcal{F}}_t^*$ is a suitable approximation to $\hat{\mathcal{F}}_t$. We counted the distribution of the error $\delta\theta$ between the real and estimated yaw angles from 378k samples. It turns out that $\delta\theta \sim \mathcal{N}(-0.4,\ 14.9^2)$ (degree), which means that 95.4% of the estimated angles have an error less than 30° (2σ rule), as depicted in Figure 4. Thus, our approximation of θ is fairly reasonable.

3.3. Ablation Study

The proposed Attention Block is designed to assign the weight of others' influences to each agent. In this experiment, we substitute a Multi-Layer Perception (MLP) Block for it and summarize the result in Table 1. Results demonstrate the superiority of the Attention Block, which captures the cross-agent relevance.

3.4. Quatitative results on INTERACTION

The INTERACTION dataset requires 3s predicted future trajectories, which is less challenging compared to nuScenes and Argoverse 2 (6s long). Thus, the demonstration of IPCC-TP's ability to model multi-agent relevance over long periods of time is limited when evaluated on INTERACTION. Regardless, we still provide our results in Table 2 below for completeness. Although the baseline method AutoBots already demonstrates satisfactory results, IPCC-TP can still exceed its performance.

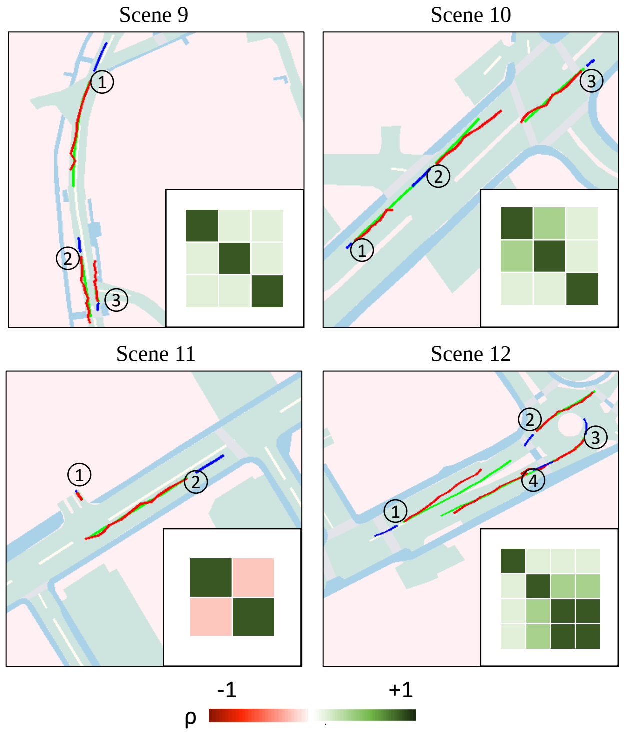

Figure 3. Visualization of IPCC matrices in nuScenes (blue line: past trajectories, green line: ground truth, red line: predictions). **Scene 9**: Vehicle ① is far away from vehicle ② and ③, and the latter two vehicles are driving in opposite directions. **Scene 10**: Vehicle ① is following vehicle ②, and vehicle ③ is driving on the opposite lane. **Scene 11**: Vehicle ① is yielding to vehicle ②. **Scene 12**: Vehicle ①, ②, ③, ④ are passing or approaching a roundabout, and vehicle ③ is closely following vehicle ④.

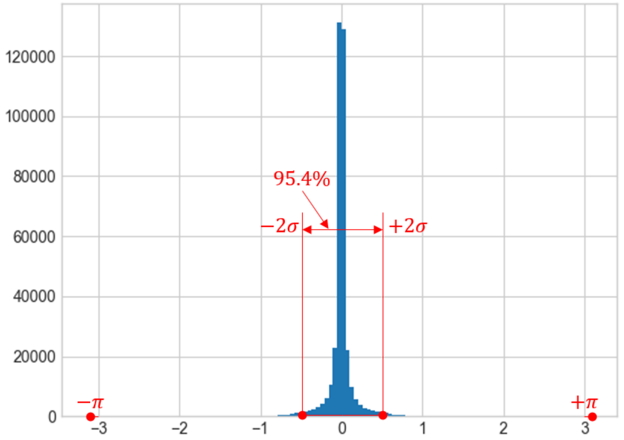

Figure 4. **Distribution of $\delta\theta$.** According to 378k samples from the nuScenes dataset, the error $\delta\theta$ between the real yaw angle ϕ and the estimated yaw angle θ has a mean value of $-0.4°$ and a standard deviation of $14.9°$.

INTERACTION	minJ-ADE(6)	minJ-FDE(6)
AutoBots	0.36	0.97
AutoBots+IPCC-TP	**0.35**	**0.93**

Table 2. Evaluation on INTERACTION.

References

[1] Interpret multi-agent prediction and conditional multi-agent prediction. https://github.com/interaction-dataset/INTERPRET_challenge_multi-agent. Accessed November 18, 2022. 1

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[3] INTERPRET Challenge. http://challenge.interaction-dataset.com, 2021. 1

[4] Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. In *International Conference on Learning Representations*, 2022. 1

[5] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021. 1

[6] 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. 1

[7] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kümmerle, Hendrik Königshof, Christoph Stiller, Arnaud de La Fortelle, and Masayoshi Tomizuka. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps. *arXiv:1910.03088 [cs, eess]*, 2019. 3