Supplementary Material
VISTA: A Vision and Intent-Aware Social Attention Framework for Multi-Agent Trajectory Prediction

Anonymous WACV Algorithms Track submission

Paper ID 955

1. Quantitative results on MADRAS dataset

We present below the detailed evaluation of our method along with the strongest baselines. The results are computed separately for each scene of the dataset.

Method	ADE/FDE	minADE/minFDE	Collision Rate
Y-Net [2]	8.74 / 15.29	0.50 / 0.65	5.36%
MART [1]	0.68 / 1.27	0.17 / 0.23	2.29%
TUTR [3]	1.24 / 2.12	0.35 / 0.54	N.A.
VISTA	0.62 / 1.12	0.18 / 0.25	0.03%

Table 1. Performance on MADRAS scene 1A (in meters).

Method	ADE/FDE	minADE/minFDE	Collision Rate
Y-Net [2]	8.74 / 15.29	0.50 / 0.65	5.36%
MART [1]	0.63 / 1.19	0.17 / 0.23	1.99%
TUTR [3]	1.36 / 2.34	0.38 / 0.62	N.A.
VISTA	0.65 / 1.15	0.18 / 0.24	0.02%

Table 2. Performance on MADRAS scene 1B (in meters).

Method	ADE/FDE	minADE/minFDE	Collision Rate
Y-Net [2]	8.74 / 15.29	0.50 / 0.65	5.36%
MART [1]	0.71 / 1.34	0.18 / 0.24	2.10%
TUTR [3]	0.68 / 1.04	0.36 / 0.53	N.A.
VISTA	0.65 / 1.15	0.18 / 0.24	0.05%

Table 3. Performance on MADRAS scene 1C (in meters).

Method	ADE/FDE	minADE/minFDE	Collision Rate
Y-Net [2]	8.74 / 15.29	0.50 / 0.65	5.36%
MART [1]	0.52 / 0.97	0.14 / 0.18	3.51%
TUTR [3]	0.77 / 0.99	0.33 / 0.45	N.A.
VISTA	0.64 / 1.14	0.17 / 0.25	0.04%

Table 4. Performance on MADRAS scene 2A (in meters).

Method	ADE/FDE	minADE/minFDE	Collision Rate
Y-Net [2]	8.74 / 15.29	0.50 / 0.65	5.36%
MART [1]	0.59 / 1.13	0.16 / 0.23	2.02%
TUTR [3]	1.11 / 1.54	0.35 / 0.48	N.A.
VISTA	0.66 / 1.13	0.18 / 0.25	0.02%

Table 5. Performance on MADRAS scene 2B (in meters).

Method	ADE/FDE	minADE/minFDE	Collision Rate
Y-Net [2]	8.74 / 15.29	0.50 / 0.65	5.36%
MART [1]	0.86 / 1.63	0.17 / 0.24	2/12%
TUTR [3]	0.98 / 1.46	0.39 / 0.59	N.A.
VISTA	0.63 / 1.12	0.18 / 0.26	0.02%

Table 6. Performance on MADRAS scene 2C (in meters).

Method	ADE/FDE	minADE/minFDE	Collision Rate
Y-Net [2]	8.74 / 15.29	0.50 / 0.65	5.36%
MART [1]	0.71 / 1.35	0.18 / 0.25	14.29%
TUTR [3]	0.73 / 1.13	0.41 / 0.61	N.A.
VISTA	0.65 / 1.14	0.18 / 0.25	0.02%

Table 7. Performance on MADRAS scene 2D (in meters).

Method	ADE/FDE	minADE/minFDE	Collision Rate
Y-Net [2]	8.74 / 15.29	0.50 / 0.65	5.36%
MART [1]	0.77 / 1.31	0.22 / 0.27	1.80%
TUTR [3]	0.70 / 1.05	0.35 / 0.55	N.A.
VISTA	0.64 / 1.13	0.19 / 0.26	0.03%

Table 8. Performance on MADRAS scene 2E (in meters).

Method	ADE/FDE	minADE/minFDE	Collision Rate
Y-Net [2]	8.74 / 15.29	0.50 / 0.65	5.36%
MART [1]	0.73 / 1.43	0.17 / 0.27	1.99%
TUTR [3]	0.60 / 1.01	0.43 / 0.71	N.A.
VISTA	0.64 / 1.13	0.18 / 0.24	0%

Table 9. Performance on MADRAS scene 2F (in meters).

2. Qualitative results on MADRAS dataset

Figure 1 accompanies the interpretation of the social attention provided in the final paragraphs of Section 4.4 in the main paper.

Figure 2 and 3 show the social attention maps and the long-term trajectory predictions in SDD dataset. This example presents a scene involving four individuals, two of whom form a group (pedestrians **0** and **1**) that remains nearly motionless throughout the sequence. Only one pedestrian is in motion (**3**), approaching another stationary pedestrian (**2**). Analysis of the attention maps indicates that pedestrians **0** and **1** focus exclusively on each other. Notably, pedestrian **1** maintains a consistent level of attention toward pedestrian **0**, whereas pedestrian **0** demonstrates increasing self-focus, which may indicate that pedestrian **1** is following pedestrian **0**. In contrast, pedestrians **2** and **3** exhibit stable mutual attention throughout the sequence. Additionally, pedestrian **3**, who moves toward pedestrian **2**, receives more attention from the stationary pedestrian than is reciprocated.

References

[1] Seongju Lee, Junseok Lee, Yeonguk Yu, Taeri Kim, and Kyoobin Lee. Mart: Multiscale relational transformer networks for multi-agent trajectory prediction. In *ECCV 2024*, pages 89–107. Springer, 2024. 1, 2

[2] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *ICCV 2021*, pages 15202–15212, 2021. 1, 2

[3] Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *ICCV 2023*, pages 9675–9684, 2023. 1, 2

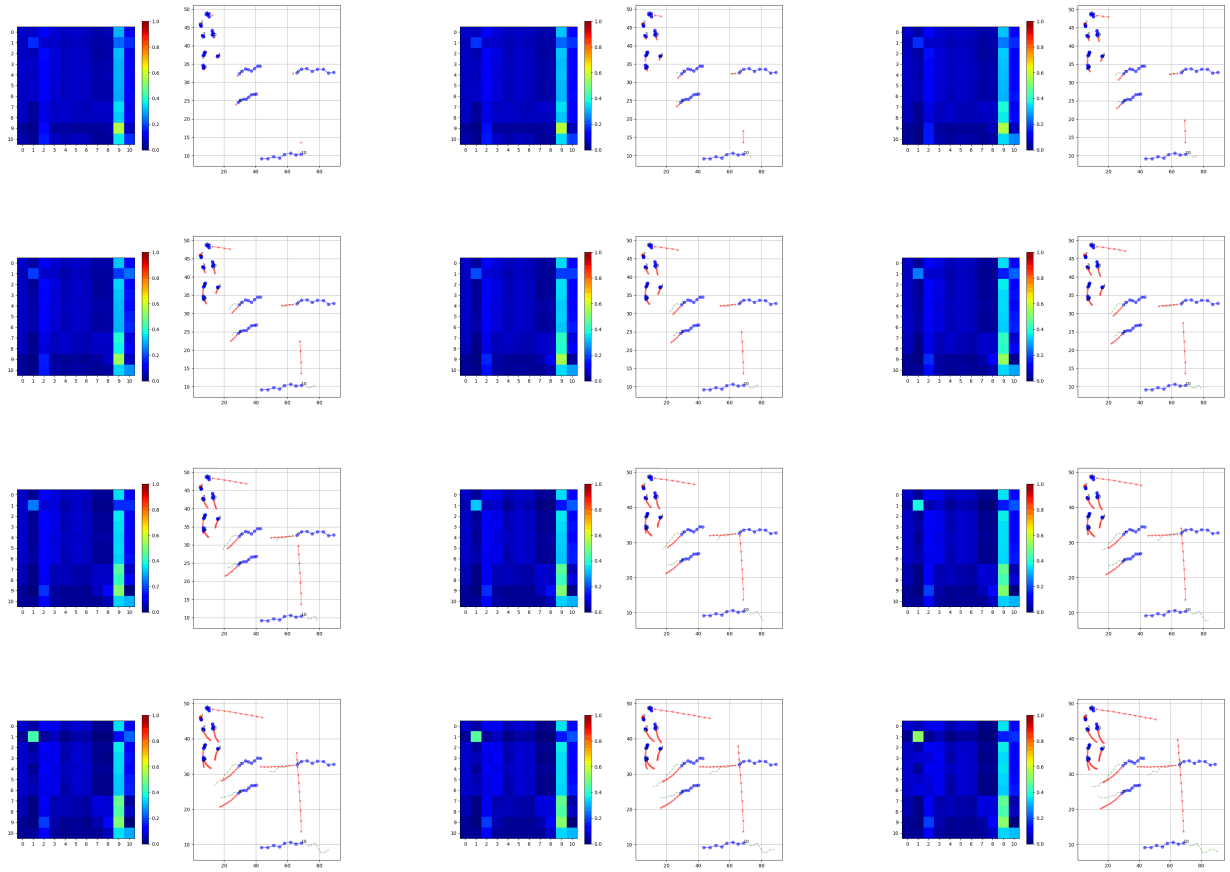

Figure 1. **Example of a prediction in MADRAS Dataset.** Each tile shows the social-attention matrix on the left and the predicted trajectories in red, compared to the future ground-truth trajectories in green on the right.

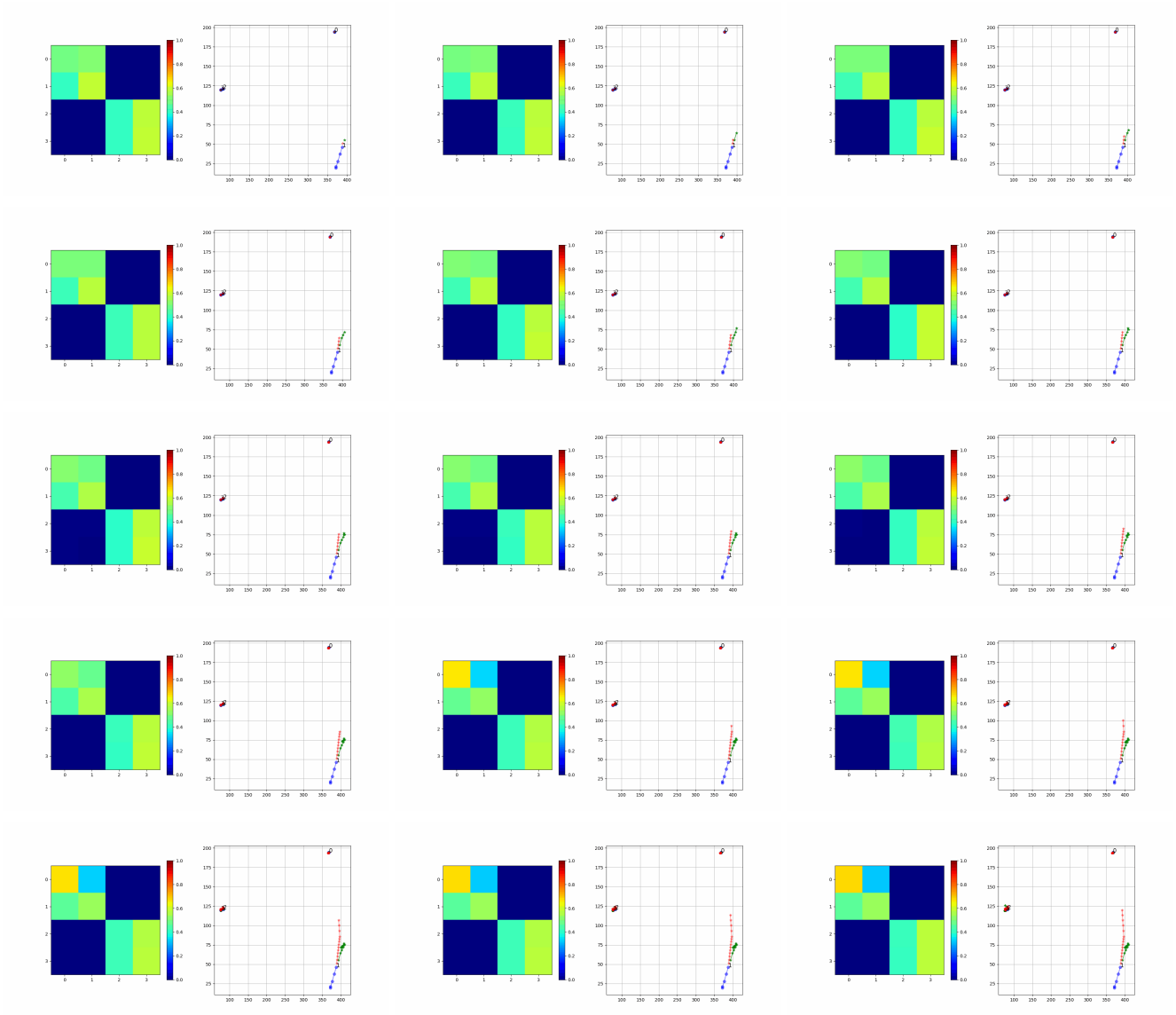

Figure 2. **Example of a prediction in SDD dataset for a long-term prediction (5+30).** This figure shows the first 15 frames of the sequence. Each tile shows the social-attention matrix on the left and the predicted trajectories in red, compared to the future ground-truth trajectories in green on the right

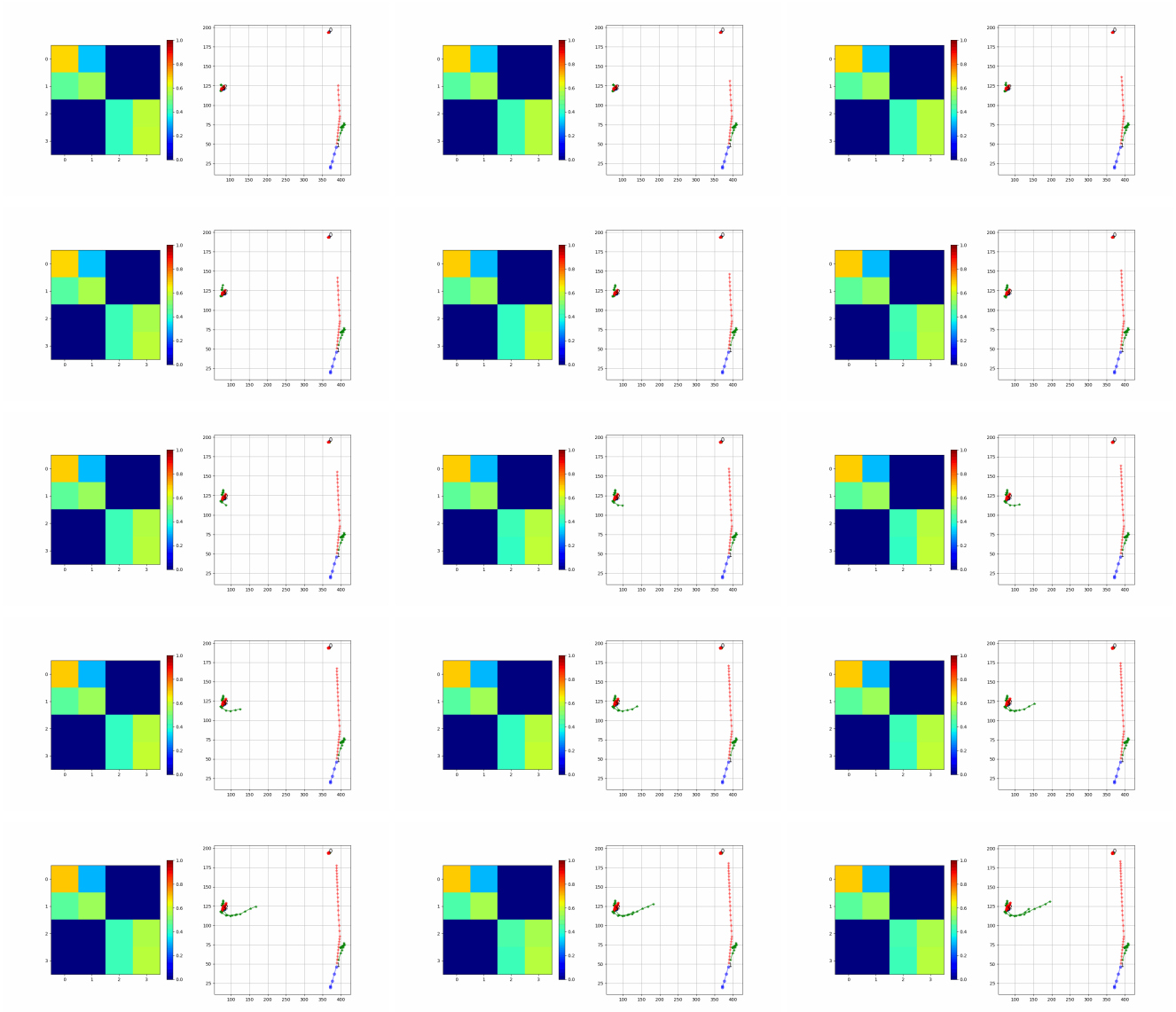

Figure 3. **Example of a prediction in SDD dataset for a long-term prediction (5+30).** This figure shows the last 15 frames of the sequence. Each tile shows the social-attention matrix on the left and the predicted trajectories in red, compared to the future ground-truth trajectories in green on the right