2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

MATRIX: Multi-Agent Trajectory Generation with Diverse Contexts

Zhuo Xu*[1], Rui Zhou*[2], Yida Yin*[3], Huidong Gao[3], Masayoshi Tomizuka[3], and Jiachen Li[4]

***Abstract*— Data-driven methods have great advantages in modeling complicated human behavioral dynamics and dealing with many human-robot interaction applications. However, collecting massive and annotated real-world human datasets has been a laborious task, especially for highly interactive scenarios. On the other hand, algorithmic data generation methods are usually limited by their model capacities, making them unable to offer realistic and diverse data needed by various application users. In this work, we study trajectory-level data generation for multi-human or human-robot interaction scenarios and propose a learning-based automatic trajectory generation model, which we call Multi-Agent TRajectory generation with dIverse conteXts (MATRIX). MATRIX is capable of generating interactive human behaviors in realistic diverse contexts. We achieve this goal by modeling the explicit and interpretable objectives so that MATRIX can generate human motions based on diverse destinations and heterogeneous behaviors. We carried out extensive comparison and ablation studies to illustrate the effectiveness of our approach across various metrics. We also presented experiments that demonstrate the capability of MATRIX to serve as data augmentation for imitation-based motion planning.**

I. Introduction

Interactive human behavioral dynamics are among the most challenging dynamics to model due to the complicated hidden features and the diverse behaviors. Researchers have recently favored data-driven methods as the solution for a wide range of human-robot interactive problems. Nevertheless, collecting real-world human motion datasets even only on the trajectory level is not an easy task because it can take huge amounts of human volunteer recruiting or label annotation efforts. Although some algorithmic methods can generate a variety of human trajectory data based on deterministic or stochastic algorithmic motion generators [1], [2], most of these rule-based approaches can only do well in limited domains and yet fail to produce realistic and smooth trajectory data for general purposes. While the demand for massive human behavior datasets is further exacerbated by the advance of deep learning-based methods, the plenty of available trajectory data can also benefit the learning of the human motion logic, which in turn assists assorted downstream tasks with human-robot interaction. To this end, equipped with deep neural networks, researchers have built a variety of generative models by encoding highly multi-modal and uncertain human motions into latent states [3], [4], [5], [6]. However, these models are based on the assumption that the underlying randomness in human behaviors is a conditional Gaussian distribution. Thus, their performances depend on how well the latent space models the hidden features of agents, and the diversity of the generated data cannot be guaranteed.

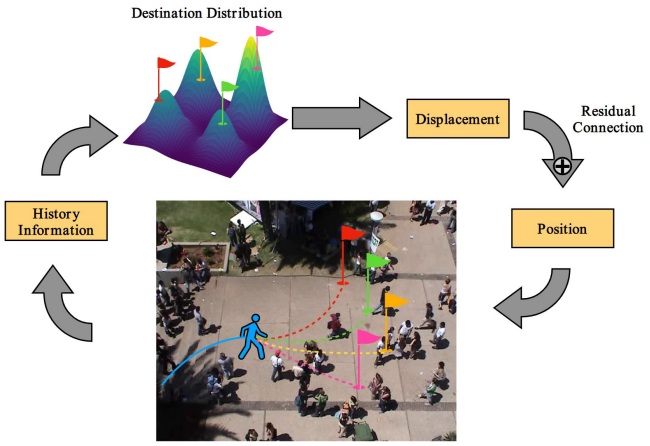

Fig. 1: **Illustration of the MATRIX generation process.** Based on the observed trajectory and sampled destinations, MATRIX can generate various heterogeneous trajectories through residual connection.

To achieve explainable learning of the human behavioral latent features, we enable the human trajectory generative model to produce diverse and distinct contexts by explicitly modeling one of the most prominent properties that affect human behaviors: the temporal traveling destinations. We call our model Multi-Agent TRajectory generation with dIverse conteXts (MATRIX), which adopts a conditional variational autoencoder framework and self-supervised training scheme while featuring a Gaussian Mixture Model (GMM) for modeling the hidden distribution of temporal destinations, which naturally exhibits multi-modality, meaning diverse future interaction modes can emerge from same past trajectory contexts. In addition, GMM provides explainable parameters that we can regularize throughout training to forfeit mode collapse and guarantee diverse generated behaviors. As a realistic human trajectory data generator, MATRIX also enforces soft safety constraints by adopting residual structures that produce human actions. The illustration of the trajectory generation process is shown in Fig. 1. We evaluated MATRIX in the human crowd navigation setting, against a variety of baseline approaches and ablation models, with a series of metrics demonstrating MATRIX's capability of generating realistic and diverse trajectory data. The main contributions of this paper are as follows:

1) We present a novel human trajectory generation frame-

*indicates equal contribution.

[1]Everyday Robots, X, USA zhuoxu@google.com
[2]Massachusetts Institute of Technology, USA zhourui@mit.edu
[3]University of California, Berkeley, USA {davidyinyida0609, hgao, tomizuka}@berkeley.edu
[4]University of California, Riverside, USA jiachen.li@ucr.edu

work called MATRIX, which produces diverse and realistic human motion data.
2) We design a GMM to explicitly model the distribution of human temporal destinations and utilize residual action to control the aggressiveness of sampled trajectories and encourage diverse generated behaviors.
3) We introduce several novel motion primitive distribution shifts (Chi-square distance χ^2 between real and generated trajectories) as the realism metrics in addition to the classic waypoint displacement error metrics.
4) Our approach, as evaluated against a series of baselines and ablations, achieves state-of-the-art performance as a diverse trajectory data generator, in terms of quantitative diversity and realism metrics on the ETH/UCY benchmarks, and as demonstrated in experiments where it serves as a data source for the training set augmentation for a downstream behavior cloning task.

II. Related Work

A. Human Trajectory Prediction and Planning

Data-driven navigation behavior, interaction understanding, prediction and planning have garnered significant attention from numerous communities in recent years [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. Some of the earlier works in human trajectory forecasting and planning include the social force model [17], the dynamic potential field [18], velocity-based collision avoidance [2], and model predictive control [19]. Machine learning has also been utilized to forecast human trajectories, by modeling it as a deterministic time-series regression problem and solving it using Gaussian Process Regression (GPR) [20], inverse reinforcement learning (IRL) [21], and recurrent neural networks (RNNs) [22], [23]. More recent generative approaches have adopted a recurrent architecture with a latent space, such as a conditional variational auto-encoder (CVAE) [24], [25], [3], [26], [27], [28], [29], [6], a generative adversarial network (GAN) [30], [31], [32], or a diffusion model [33], [34], [35], [36] to encode multi-modality. Some other works also focus on improving the stochastic process for multi-modal prediction [37]. On the modeling side, Graph Convolutional Networks (GCN) are first introduced in [38], and Spatio-Temporal Graph Convolutional Networks (STGCN) [39] and Social-STGNN [40] are designed to capture both spatial and temporal information.

There are also methods for human trajectory prediction that employ attention mechanisms [29], [41], [42], [33], [43], [44]. AgentFormer [41] employs an agent-aware attention mechanism. MID [33] is a Transformer-based framework with motion indeterminacy diffusion. ScePT [45] generates scene-consistent joint trajectory predictions with a tunable risk measure. Y-net [46] utilizes encoder and decoder architecture to reconstruct the heatmap of future trajectories. [47] introduces a socially attentive network that consists of an interactive module that encodes interactions through local maps. [48] uses model-based deep reinforcement learning to plan actions.

Compared to previous work, our method has better performance when evaluated as a data generator. The diverse trajectories generated by MATRIX show its potential for the downstream task and as an alternative to traditional data augmentation methods, such as translating and rotating.

B. Multi-modality Encoding

A key in predicting human behavior is encoding the multi-modality nature [33], [49]. Many works model agents' future modes implicitly as latent variables, including DESIRE [27] utilizing a conditional variational auto-encoder to obtain a diverse set of future prediction samples. PRECOG [50] uses a flow-based generative model to perform both standard forecasting and the novel task of conditional forecasting. SocialGAN [30] further trains a network adversarially against a recurrent discriminator, and encourage diverse predictions with a novel variety loss. Some other works choose to discretize the output space as goals or anchors, then do predictions based on each goal or anchor, including MultiPath [51] and Covernet [52]. TNT [53] and DenseTNT [54] also first predict goal candidates and then generate separate trajectories conditioned on targets. In comparison, we utilize a Gaussian Mixture Model (GMM) to capture the goal of continuous distribution. This reduces computation load and improves the performance of goal estimation without relying on the quality of predefined goal anchors.

III. MATRIX

In this section, we introduce our approach MATRIX with a focus on encouraging the generation of diverse trajectories. At each time step t, we have N pedestrians. The 2D position of pedestrian i at time step t is denoted as $s_i^t \in \mathbb{R}^2$. We then can represent the past trajectories for all pedestrians in the past H time steps as $x = s_{1,..,N}^{t-H+1:t} \in \mathbb{R}^{N\times H\times 2}$ and the respective future trajectories in the next F time steps as $y = s_{1,..,N}^{t+1:t+F} \in \mathbb{R}^{N\times F\times 2}$. MATRIX is trained to model the distribution of $\mathbb{P}(y \mid x)$. To achieve the objective of realism, the behaviors generated by MATRIX shall match that of the real human behaviors recorded in the social pedestrian trajectory datasets. In the following, we present how MATRIX takes into consideration real data matching as well as the inductive biases for multi-modal properties. Fig. 2 illustrates the full architecture of MATRIX.

A. History-and-Interaction-Aware Context Extraction

To encode the highly dynamic and interactive social navigation scene, we adapt the encoder of Trajectron++ [3], including one node history LSTM and one edge history LSTM, denoted as f, to extract the rich and interactive context information from the multi-agent scene. We then represent the encoded history of all past trajectories as $e = f(x)$.

B. Explicit Latent Variable Modeling for Multi-Future Destination Reasoning

Temporal traveling destination is one of the most significant factors that affect human behavior. However, although in

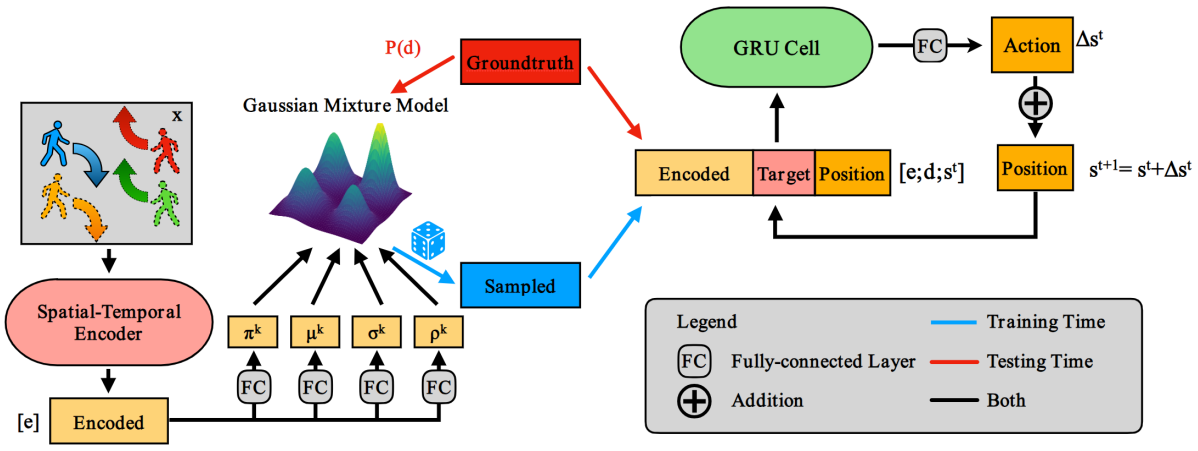

Fig. 2: **The architecture of MATRIX.** MATRIX consists of a Spatial-Temporal Encoder, a Gaussian Mixture Model (GMM), and a Gated Recurrent Unit (GRU) Decoder with a residual layer.

real crowd navigation scenes one can query the ground-truth future position as the temporal traveling destination, in data generation, the human shall exhibit multi-modal behavior. Therefore, in MATRIX, the first inference stream is to use an explicit latent stochastic model to capture this property.

Since the human potential destinations naturally follow the multi-kernel pattern, we assume joint normal distribution and design a two-dimensional Gaussian mixture model (GMM) to capture the temporal destination distribution. Concretely, the probability density function of temporal destination for each agent i is $d_i \in \mathbb{R}^2$ is written as

$$P(d_i) = \sum_{k=1}^{K} c_i^k \cdot \frac{e^{-\frac{1}{2}(d_i-\mu_i^k)(\Sigma_i^k)^{-1}(d_i-\mu_i^k)}}{2\pi \cdot \sqrt{|\Sigma_i^k|}}, \tag{1}$$

where $c_i^k \in \mathbb{R}, \mu_i^k \in \mathbb{R}^2, \Sigma_i^k \in \mathbb{R}^{2\times 2}$ are the weight, mean, and covariance matrices for the kth normal distribution and K is the number of Gaussian kernels. To obtain each of these quantities, we parameterize the encoded history with four fully connected layers $\{g_i\}_{i=1,2,3,4}$, each of which outputs the weights, means, log variances, and correlations of Gaussian kernels. In other words, we can rewrite the temporal destination inference stream as

$$\begin{aligned} \pi_i^{1:K} &= g_1(e), \mu_i^{1:K} = g_2(e), \\ \sigma_i^{1:K} &= \exp\left(g_3(e)\right), \rho_i^{1:K} = \tanh\left(g_4(e)\right). \end{aligned} \tag{2}$$

We then reconstruct the covariance matrix of the kth component from the variance σ_i^k and the correlation ρ_i^k by

$$\begin{aligned} \begin{bmatrix} \sigma_x \\ \sigma_y \end{bmatrix} &= \sigma_i^k, \rho_{xy} = \rho_i^k, \\ \Sigma_i^k &= \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}. \end{aligned} \tag{3}$$

We select K to be large enough to capture the multi-modality. To encourage MATRIX to generate trajectories that match the real data, we employ an objective of maximum log-likelihood for the ground-truth temporal target $\bar{d}_i$ in the training objective:

$$\mathscr{L}_{\text{destination}} = -\frac{1}{N}\sum_{i=1}^{N} \log P(\bar{d}_i). \tag{4}$$

In practice, since the training data is sparse and cannot represent the diverse potential future that MATRIX should be able to generate, the learned GMM can suffer from the mode collapse problem. We therefore apply a regularization on the weight, variance of the kernels, and distances between their centers. The resulting auxiliary loss is the summation of a series of hinge losses, which is written as

$$\begin{aligned} \mathscr{L}_{\text{mode\_collapse}} = &\sum_{i=1}^{N}\sum_{k_1\neq k_2} \alpha_1 h(1-\beta_1\|\mu_i^{k_1}-\mu_i^{k_2}\|_2) \\ &+ \sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_2 h(\beta_2 c_i^k - 1) + \alpha_3 h(\beta_3\|\sigma_i^k\|_2 - 1), \end{aligned} \tag{5}$$

where α's and β's are hyperparameters and $h(\cdot)$ is the hinge loss function defined as

$$h(x) = \max(0, x). \tag{6}$$

In the training time, we use the final position of the ground-truth trajectory $\bar{d}_i$ for the downstream decoder output. In the testing time, we sample diverse destinations $\hat{d}_i$ from GMMs and generate corresponding future motions with the decoder.

C. Generation of Residual Actions

After obtaining the latent representation of the spatio-temporal graph and temporal destination, the future motion is inferred using a Gated Recurrent Unit (GRU) decoder. The GRU cell is denoted as $g(\cdot)$ and the residual layer is denoted as $q(\cdot)$. Through the residual connection, MATRIX could autoregressively output the predicted trajectory by

$$\begin{aligned} h_i^t &= \begin{cases} g([e;\bar{d}_i;\hat{s}_i^t]) & \text{in training} \\ g([e;\hat{d}_i;\hat{s}_i^t]) & \text{in testing} \end{cases} \\ \Delta\hat{s}_i^t &= q(h_i^t), \hat{s}_i^{t+1} = \hat{s}_i^t + \Delta\hat{s}_i^t, \end{aligned} \tag{7}$$

where e is the encoded history of all the past trajectories, $\bar{d}_i$ is the ground truth temporal target, $\hat{d}_i$ is the sampled temporal destination from the GMM, h_i^t is the hidden state of GRU at time step t, $\Delta\hat{s}_i^t$ is the residual displacement, and $\hat{s}_i^t$ is the predicted state for agent i at time step t.

TABLE I: **ASD/ADE/FDE values of 20 samples on the ETH/UCY dataset.** Bold indicates best.

	UNIV			HOTEL			ZARA1			ZARA2			ETH		
	ASD	**ADE**	**FDE**	**ASD**	**ADE**	**FDE**	**ASD**	**ADE**	**FDE**	**ASD**	**ADE**	**FDE**	**ASD**	**ADE**	**FDE**
MID[33]	N/A	0.22	0.45	N/A	0.13	0.22	N/A	0.17	0.30	N/A	0.13	0.27	N/A	0.39	0.66
PECNet[55]	N/A	0.35	0.60	N/A	0.18	0.24	N/A	0.22	0.39	N/A	0.17	0.30	N/A	0.54	0.87
Y-Net[56]	N/A	0.24	0.41	N/A	0.10	**0.14**	N/A	0.17	**0.27**	N/A	0.13	**0.22**	N/A	**0.28**	**0.33**
Trajectron++[3]	1.38	0.22	0.42	1.12	0.12	0.19	1.31	0.17	0.32	1.08	**0.12**	0.25	1.71	0.44	0.85
AgentFormer[41]	0.13	0.25	0.45	1.00	0.14	0.22	0.68	0.18	0.30	0.44	0.14	0.24	2.76	0.45	0.74
Social Implicit[40]	1.26	0.31	0.60	2.39	0.20	0.36	1.08	0.26	0.51	1.20	0.22	0.43	1.30	0.67	1.47
ExpertTraj+GMM[57]	0.21	**0.19**	0.44	0.11	**0.09**	0.15	0.21	**0.15**	0.31	0.14	**0.12**	0.24	0.48	**0.30**	0.62
MATRIX	**2.72**	0.22	**0.39**	**2.87**	0.19	0.29	**2.86**	0.20	0.35	**2.41**	0.15	0.27	**3.27**	0.94	1.61

TABLE II: χ^2 **distances on the ETH/UCY dataset.** Bold indicates best.

Generated Data	Velocity	Acceleration	Angular Velocity	Angular Acceleration
Imitation Learning Data	1.740	2.615	0.103	0.004
Trajectron++ Data	**0.129**	0.889	0.071	**0.002**
Agentformer Data	0.645	1.226	0.025	0.008
MATRIX Data	0.184	**0.763**	**0.016**	**0.002**

To ensure MATRIX predicts correct motions, the sequential network is optimized by minimizing the Huber error between the predicted trajectory $\hat{s}_i^\tau$ and the ground-truth trajectory s_i^τ. The reconstruction objective is written as

$$\mathscr{L}_{\text{reconstruction}} = \frac{1}{NF}\sum_{i=1}^{N}\sum_{\tau=t+1}^{t+F} L_{\text{Huber}}(\hat{s}_i^\tau, s_i^\tau). \quad (8)$$

Overall, we train the network to minimize the combined loss

$$\mathscr{L} = \lambda_1\mathscr{L}_{\text{destination}} + \lambda_2\mathscr{L}_{\text{mode\_collapse}} + \lambda_3\mathscr{L}_{\text{reconstruction}}, \quad (9)$$

where λ_1, λ_2, and λ_3 are hyperparameters.

IV. Experimental Evaluation

To demonstrate the effectiveness of MATRIX, we first illustrate the setup of the experiments. Then we describe a series of metrics that we selected and designed to evaluate the performance of MATRIX both as a predictor and a generator. Finally, we train an imitation-based motion planner on synthetic data generated by MATRIX and demonstrate the diversity of our generated results as well as the regularization effects of MATRIX data.

A. Experiment Setup

1) Data Preparation: MATRIX is trained on two widely used datasets: the ETH dataset [58], with subsets named ETH and HOTEL, and the UCY dataset [59], with subsets named ZARA1, ZARA2, and UNIV. Both datasets provide interactive human pedestrian navigation episodes and provide key information on pedestrian position and velocity. In our experiments, we use the real initial state from the dataset as initialization. In the training phase, the data is segmented into batches that consist of observed trajectories of 8 time steps, each of which corresponds to 3.2 seconds, and future trajectories of the next 12 time steps, which corresponds to 4.8 seconds. We follow the leave one out cross-validation as previous works [30], [3].

2) Implementation Details: We designed four fully connected layers to model the weights, means, variances, and covariances of GMM. We set K the number of GMM components to 4. For the decoder, we employ a Gated Recurrent Unit (GRU), whose hidden size is 128, and a linear layer to output the residual action. MATRIX is implemented with PyTorch and trained with Intel Core I7 CPUs and NVIDIA RTX 2080 Ti GPUs for 100 epochs. The training iterations take a data batch of size 256. The learning rate is set to 0.001 initially and decays exponentially every epoch with a decay rate of 0.9999. The model is trained with Adam optimizer and gradients are clipped at 1.0.

3) Evaluation Metrics: We compare MATRIX against baselines based on a series of metrics demonstrating the diversity and realism properties of the generated trajectories. Above all, we evaluate how well MATRIX produces diverse and realistic contexts. Therefore, we include:

a) Diversity (ASD): The diversity of the generated data is an important metric. We adopt the average self distance (ASD) [60] to measure how the generator can produce diverse contexts. To compute the ASD, we generate samples and filter out the trajectories with any collision, resulting in 20 samples. Then, we find the maximum of the average distances across time between any of the two samples $s_{l_1,i}^\tau$ and $s_{l_2,i}^\tau$:

$$\text{ASD} = \frac{1}{F}\max_{l_1,l_2\in\{1,\cdots,20\}}\sum_{\tau=t+1}^{t+F}\|s_{l_1,i}^\tau - s_{l_2,i}^\tau\|_2, \quad (10)$$

We follow previous trajectory reconstruction literature and report the distance error metrics as a proxy of our model's capability of producing realistic trajectories [61], [3]:

b) Average Displacement Error (ADE): ADE is the mean l_2 distance between the ground truth s_i^τ and predictions $\hat{s}_i^\tau$:

$$\text{ADE} = \frac{1}{N}\frac{1}{F}\sum_{i=1}^{N}\sum_{\tau=t+1}^{t+F}\|\hat{s}_i^\tau - s_i^\tau\|_2. \quad (11)$$

c) Final Displacement Error (FDE): FDE is the l_2 distance between the predicted final position s_i^{t+F} and the

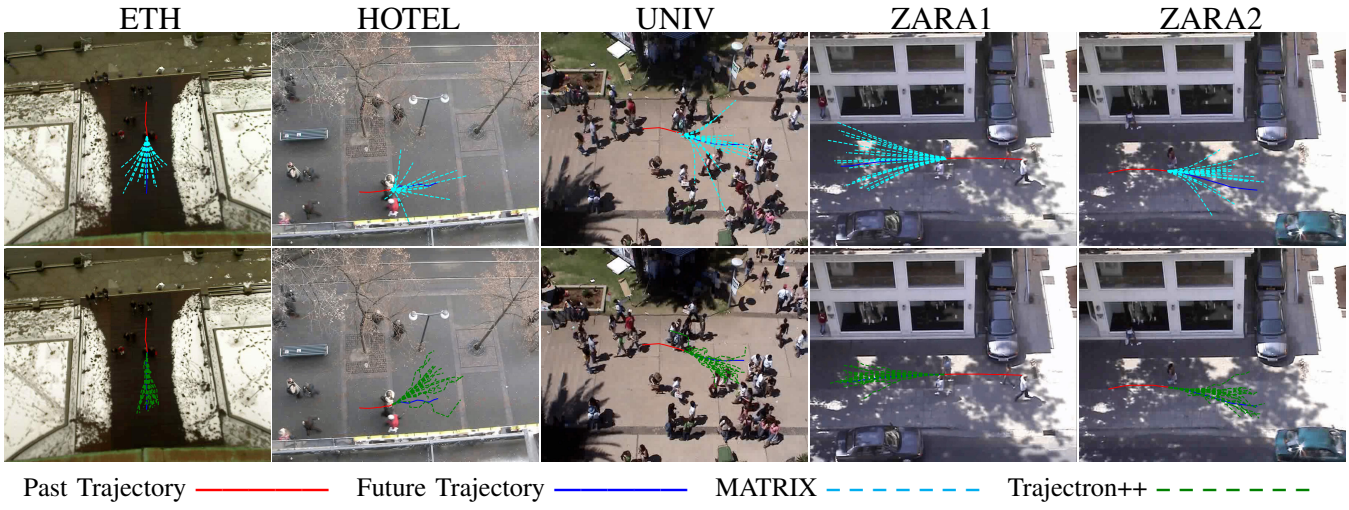

Fig. 3: **Visualization of generated trajectories.** Provided with the past trajectory (red), MATRIX (cyan) and Trajectron++ (green) can generate 20 possible future trajectories for five different scenes. We see that our generated trajectories are much more diverse than Trajectron++. Zoom in for better visualization.

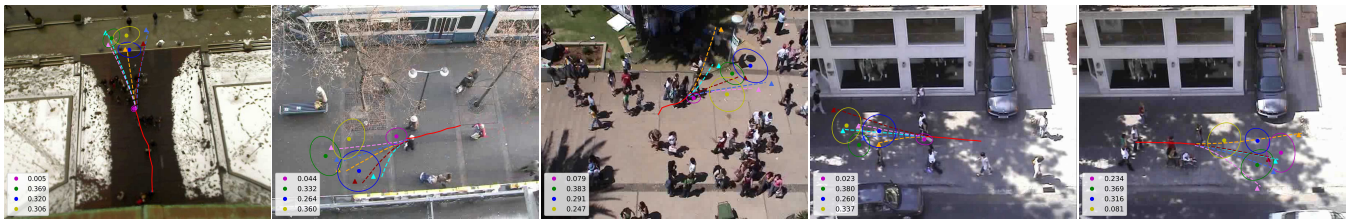

Fig. 4: **Visualization of GMM.** We use MATRIX to generate five future trajectories (orange, cyan, violet, dark red, and royal blue) based on the past trajectory (red) and five stochastic destinations, represented as triangles, sampled from GMM. The center of each ellipse (green, magenta, yellow, and dark blue) is the mean of each Gaussian, and the radius is its one standard deviation. The weight of each Gaussian can be found in the legend. Zoom in for better visualization.

ground-truth final position $\hat{s}_i^{t+F}$:

$$\text{FDE} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{s}_i^{t+F} - s_i^{t+F}\|_2, \tag{12}$$

where N is the total number of pedestrians and F is the number of future horizons.

Since MATRIX is supposed to generate stochastic multi-modal contexts, we sample 20 trajectories and evaluate the best matching mode, and compute the best-of-20 ADE/FDE.

d) Chi-square Distance (χ^2): The capability of reconstructing the training trajectories is no longer a fair metric for realism under the setting of trajectory generation, and thus we introduce a new set of measurements of the Chi-square distance between key motion primitive distributions of the generated data and the original data. We compare four primitives, including velocity, acceleration, angular velocity, and angular acceleration.

$$\chi^2 = \sum_{i=1}^{20} \frac{(x_i - y_i)^2}{x_i + y_i}, \tag{13}$$

where x_i and y_i are the estimated probability density of the generated data and the raw data in the ith bin in a total of 20 bins. Since all five subsets come from similar scenarios, we compute the average Chi-square Distance over them.

B. Quantitative Comparisons

We compare our model with a wide range of state-of-the-art models in Table I. Note that some of the models, such as Y-Net and ExpertTraj-GMM, achieve lower reconstruction errors at the cost of the diversity metric ASD. In contrast, MATRIX has a significantly higher ASD value. For example, the HOTEL dataset has an ASD score of 2.87, which is 100.7% better than that of Trajectron++. Meanwhile, trajectories generated by MATRIX have ADE and FDE values with the same orders of magnitude compared with models specialized for trajectory reconstruction, demonstrating that MATRIX can maintain a relatively low reconstruction error even when the generated behaviors are far more diverse.

Since reconstructing the training trajectories is not sufficient for realism evaluation under the circumstance of diverse trajectory generation. In Table II and Fig. 5, we show how the distribution shifts of key motion primitives for trajectories generated by MATRIX compare with trajectories generated using other methods. MATRIX has the lowest χ^2 scores across most motion primitives, indicating that the MATRIX data indeed matches the distribution of human motions. This is attributed to the learnable residual action, as in contrast, Trajectron++ has unlearnable dynamic integration via dynamics, making it much more difficult to control over the randomness of GMM and produce the distributions matching the real one [3].

TABLE III: **ASD/ADE/FDE values for ablation study.** MC = Mode Collapse, $\bigoplus$ = Residual Actions. Bold indicates best.

MC	$\bigoplus$	UNIV			HOTEL			ZARA1			ZARA2			ETH		
		ASD	**ADE**	**FDE**	**ASD**	**ADE**	**FDE**	**ASD**	**ADE**	**FDE**	**ASD**	**ADE**	**FDE**	**ASD**	**ADE**	**FDE**
-	-	2.12	0.29	0.53	2.56	0.24	0.41	2.34	0.25	0.47	1.82	0.20	0.36	3.09	0.96	1.71
-	✓	2.06	0.28	0.53	2.34	0.23	0.39	2.30	0.24	0.47	1.85	0.19	0.37	2.86	0.96	1.70
✓	-	2.34	0.28	0.52	2.48	0.24	0.40	2.50	0.26	0.48	2.03	0.20	0.37	2.78	1.09	1.89
✓	✓	**2.72**	**0.22**	**0.39**	**2.87**	**0.19**	**0.29**	**2.86**	**0.20**	**0.35**	**2.41**	**0.15**	**0.27**	**3.27**	**0.94**	**1.61**

TABLE IV: **ADE/FDE values for imitation learning with data augmentations.** Bold indicates best.

	UNIV		HOTEL		ZARA1		ZARA2		ETH	
	ADE	**FDE**	**ADE**	**FDE**	**ADE**	**FDE**	**ADE**	**FDE**	**ADE**	**FDE**
Raw Data	0.38	**0.43**	0.51	0.52	0.27	0.36	**0.21**	0.28	**0.69**	**0.85**
Trajectron++ Data	0.35	**0.43**	0.46	0.70	**0.25**	**0.35**	**0.21**	0.30	0.75	1.05
MATRIX Data	**0.34**	**0.43**	**0.44**	**0.48**	**0.25**	**0.35**	**0.21**	**0.27**	0.71	0.91

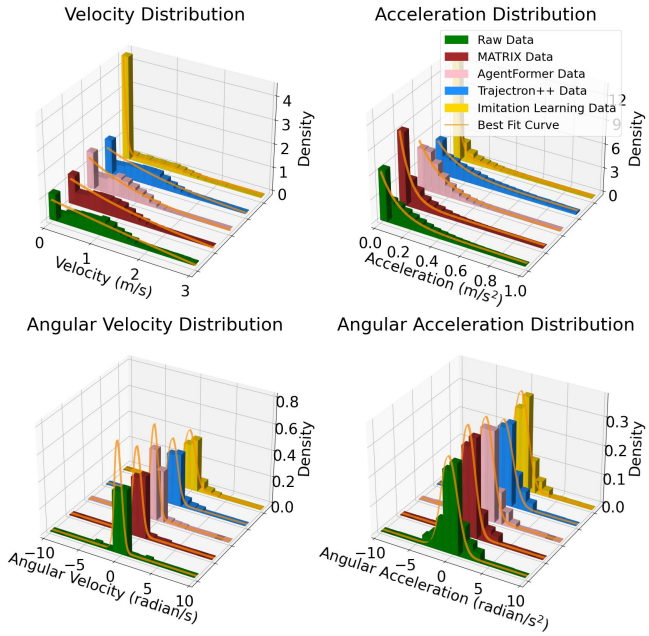

Fig. 5: **Physics primitives of the generated data.** We plot the histogram of physics primitives of four generated datasets – MATRIX data (red), Agentformer Data (pink), Trajectron++ Data (blue), and Imitation Learning Data (yellow) – against the raw one (green). The orange line is the best-fit curve. Note that we use exponential distribution for velocity and acceleration and Gaussian distribution for angular velocity and angular acceleration.

C. Multi-modality

We further visualize the diverse trajectories generated by MATRIX. Fig. 3 illustrates the 20 sampled predictions on all five subsets. The qualitative results show that though both MATRIX and Trajectron++ [3] can generate heterogeneous sequences of paths, MATRIX produces much more diverse outcomes. In addition, we observe that MATRIX's samples are much smoother than Trajectron++ and match the real behaviors of human motions. We deduce this feature is attributed to the advantage of employing both GMM and residual actions. Specifically, residual connection encourages each agent to move toward the stochastic destinations sampled from GMM and thus results in reasonable paths. To better understand how each of the two components works, we sample five different targets from GMM and visualize each respective trajectory in Fig. 4. We can see each generated trajectory is exactly driven by the target, showing the effectiveness of using GMMs to model the explicit latent variable and residual action to control the direction.

D. Ablation Study

To demonstrate the significance of each component in MATRIX, we perform a comprehensive ablation study in Table III. We show that without the mode collapse loss, MATRIX suffers from higher reconstruction errors and lower diversity because each GMM collapses to a single point with unknown variance. In addition, removing the residual action scheme leads to an increment in both displacement errors.

E. Serving as a Data Augmentation

To further investigate the realism of the data generated by MATRIX, we use the samples generated by MATRIX as an augmentation dataset. We combine UCY and ETH datasets with synthetic data generated by MATRIX from the original datasets and use the combined dataset to train imitation learning planners. We evaluate the planner performance on the evaluation datasets against planners trained using the original datasets only and augmented with synthetic data generated by Trajectron++. All models were of the same structure of 10-layer and 128 hidden-size LSTM with residual layers, and trained for 170 epochs. The results show that the imitation learning model learned using data generated by MATRIX can produce lower reconstruction errors in Table IV, which is significant since with zero extra data beyond the training data, MATRIX's generated data improved the imitation learning planner performance on unseen evaluation datasets. Hence, we conclude MATRIX can produce both diverse and realistic samples that are beneficial for downstream tasks.

V. Conclusions

In this paper, we introduce MATRIX, a data generator for multi-agent human trajectory generation with diverse contexts. By explicitly modeling the significant factor that affects heterogeneous human behaviors - the temporal destination - and controlling moving direction through a residual network, MATRIX generates multi-modal behaviors that realistically interact with external agents. Our experiments demonstrate the realism and diversity of the MATRIX data, as well as its potential to serve as a predictor.

REFERENCES

[1] J. Van den Berg, M. Lin, and D. Manocha, "Reciprocal velocity obstacles for real-time multi-agent navigation," in *2008 IEEE International Conference on Robotics and Automation*. IEEE, 2008, pp. 1928–1935.

[2] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics Research: The 14th International Symposium ISRR*. Springer, 2011, pp. 3–19.

[3] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. 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 16*. Springer, 2020, pp. 683–700.

[4] C. Choi, J. H. Choi, J. Li, and S. Malla, "Shared cross-modal trajectory prediction for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 244–253.

[5] R. Zhou, H. Zhou, H. Gao, M. Tomizuka, J. Li, and Z. Xu, "Grouptron: Dynamic multi-scale graph convolutional networks for group-aware dense crowd trajectory forecasting," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 805–811.

[6] V. M. Dax, J. Li, E. Sachdeva, N. Agarwal, and M. J. Kochenderfer, "Disentangled neural relational inference for interpretable motion prediction," *IEEE Robotics and Automation Letters*, 2023.

[7] K. Li, Y. Chen, M. Shan, J. Li, S. Worrall, and E. Nebot, "Game theory-based simultaneous prediction and planning for autonomous vehicle navigation in crowded environments," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 2977–2984.

[8] H. Chang, Z. Xu, and M. Tomizuka, "Cascade attribute network: Decomposing reinforcement learning control policies using hierarchical neural networks," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 8181–8186, 2020.

[9] Z. Xu, H. Chang, C. Tang, C. Liu, and M. Tomizuka, "Toward modularization of neural network autonomous driving policy using parallel attribute networks," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1400–1407.

[10] Z. Xu, J. Chen, and M. Tomizuka, "Guided policy search model-based reinforcement learning for urban autonomous driving," *arXiv preprint arXiv:2005.03076*, 2020.

[11] J. Chen, Z. Xu, and M. Tomizuka, "End-to-end autonomous driving perception with sequential latent representation learning," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 1999–2006.

[12] J. Li, F. Yang, M. Tomizuka, and C. Choi, "Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[13] Z. Xu and M. Tomizuka, "History encoding representation design for human intention inference," *arXiv preprint arXiv:2106.02222*, 2021.

[14] J. Li, C. Hua, H. Ma, J. Park, V. Dax, and M. J. Kochenderfer, "Multi-agent dynamic relational reasoning for social robot navigation," *arXiv preprint arXiv:2401.12275*, 2024.

[15] K. Mahadevan, J. Chien, N. Brown, Z. Xu, C. Parada, F. Xia, A. Zeng, L. Takayama, and D. Sadigh, "Generative expressive robot behaviors using large language models," *arXiv preprint arXiv:2401.14673*, 2024.

[16] S. Nasiriany, F. Xia, W. Yu, T. Xiao, J. Liang, I. Dasgupta, A. Xie, D. Driess, A. Wahid, Z. Xu, *et al.*, "Pivot: Iterative visual prompting elicits actionable knowledge for vlms," *arXiv preprint arXiv:2402.07872*, 2024.

[17] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[18] A. Treuille, S. Cooper, and Z. Popović, "Continuum crowds," *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 1160–1168, 2006.

[19] L. Sun, P.-Y. Hung, C. Wang, M. Tomizuka, and Z. Xu, "Distributed multi-agent interaction generation with imagined potential games," *arXiv preprint arXiv:2310.01614*, 2023.

[20] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer school on machine learning*. Springer, 2003, pp. 63–71.

[21] N. Lee and K. M. Kitani, "Predicting wide receiver trajectories in american football," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[22] J. Morton, T. A. Wheeler, and M. J. Kochenderfer, "Analysis of recurrent neural networks for probabilistic modeling of driver behavior," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 5, pp. 1289–1298, 2016.

[23] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *2018 IEEE international Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4601–4607.

[24] N. Deo and M. M. Trivedi, "Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1179–1184.

[25] J. Li, H. Ma, and M. Tomizuka, "Conditional generative neural system for probabilistic trajectory prediction," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6150–6156.

[26] H. Girase, H. Gang, S. Malla, J. Li, A. Kanehara, K. Mangalam, and C. Choi, "Loki: Long term and key intentions for trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9803–9812.

[27] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 336–345.

[28] G. Chen, J. Li, J. Lu, and J. Zhou, "Human trajectory prediction via counterfactual analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9824–9833.

[29] J. Li, H. Ma, Z. Zhang, J. Li, and M. Tomizuka, "Spatio-temporal graph dual-attention network for multi-agent prediction and tracking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 556–10 569, 2021.

[30] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2255–2264.

[31] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1349–1358.

[32] J. Li, H. Ma, and M. Tomizuka, "Interaction-aware multi-agent tracking and probabilistic behavior prediction via adversarial learning," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 6658–6664.

[33] T. Gu, G. Chen, J. Li, C. Lin, Y. Rao, J. Zhou, and J. Lu, "Stochastic trajectory prediction via motion indeterminacy diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 113–17 122.

[34] C. Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, D. Anguelov, *et al.*, "Motiondiffuser: Controllable multi-agent motion prediction using diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9644–9653.

[35] G. Barquero, S. Escalera, and C. Palmero, "Belfusion: Latent diffusion for behavior-driven human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2317–2327.

[36] W. Mao, C. Xu, Q. Zhu, S. Chen, and Y. Wang, "Leapfrog diffusion model for stochastic trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5517–5526.

[37] I. Bae, J.-H. Park, and H.-G. Jeon, "Non-probability sampling network for stochastic human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6477–6487.

[38] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2016.

[39] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[40] A. Mohamed, K. Qian, M. Elhoseiny, and C. 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*, 2020, pp. 14 424–14 432.

[41] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9813–9823.

[42] F.-Y. Sun, I. Kauvar, R. Zhang, J. Li, M. J. Kochenderfer, J. Wu, and N. Haber, "Interaction modeling with multiplex attention," *Advances*

in Neural Information Processing Systems, vol. 35, pp. 20 038–20 050, 2022.

[43] J. Li, F. Yang, H. Ma, S. Malla, M. Tomizuka, and C. Choi, "Rain: Reinforced hybrid attention inference network for motion forecasting," in *International Conference on Computer Vision (ICCV)*, 2021.

[44] H. Hu, Q. Wang, Z. Zhang, Z. Li, and Z. Gao, "Holistic transformer: A joint neural network for trajectory prediction and decision-making of autonomous vehicles," *Pattern Recognition*, vol. 141, p. 109592, 2023.

[45] Y. Chen, B. Ivanovic, and M. Pavone, "Scept: Scene-consistent, policy-based trajectory predictions for planning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 103–17 112.

[46] K. Mangalam, Y. An, H. Girase, and J. Malik, "From goals, waypoints & paths to long term human trajectory forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 233–15 242.

[47] C. Chen, Y. Liu, S. Kreiss, and A. Alahi, "Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6015–6022.

[48] C. Chen, S. Hu, P. Nikdel, G. Mori, and M. Savva, "Relational graph learning for crowd navigation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 007–10 013.

[49] H. Ma, J. Li, R. Hosseini, M. Tomizuka, and C. Choi, "Multi-objective diverse human motion prediction with knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8161–8171.

[50] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine, "Precog: Prediction conditioned on goals in visual multi-agent settings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2821–2830.

[51] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," *arXiv preprint arXiv:1910.05449*, 2019.

[52] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff, "Covernet: Multimodal behavior prediction using trajectory sets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 074–14 083.

[53] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid, *et al.*, "Tnt: Target-driven trajectory prediction," in *Conference on Robot Learning*. PMLR, 2021, pp. 895–904.

[54] J. Gu, C. Sun, and H. Zhao, "Densetnt: End-to-end trajectory prediction from dense goal sets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 303–15 312.

[55] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020.

[56] K. Mangalam, Y. An, H. Girase, and J. Malik, "From goals, waypoints & paths to long term human trajectory forecasting," in *Proc. International Conference on Computer Vision (ICCV)*, Oct. 2021.

[57] Z. He and R. P. Wildes, "Where are you heading? dynamic trajectory prediction with expert goal examples," in *Proceedings of the International Conference on Computer Vision (ICCV)*, Oct. 2021.

[58] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 261–268.

[59] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Computer graphics forum*, vol. 26, no. 3. Wiley Online Library, 2007, pp. 655–664.

[60] Y. Yuan and K. M. Kitani, "Diverse trajectory forecasting with determinantal point processes," in *International Conference on Learning Representations*, 2019.

[61] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.