JACoP: Joint Alignment for Compliant Multi-Agent Prediction

Qingze (Tony) Liu[1] Alen Mrdovic[1] Danrui Li[1] Mathew Schwartz[1]
Sejong Yoon[2] Mubbasir Kapadia[1] Vladimir Pavlovic[1]
[1]Rutgers University, New Brunswick
[2]The College of New Jersey

Abstract

Stochastic Human Trajectory Prediction (HTP) using generative modeling has emerged as a significant area of research. Although state-of-the-art models excel in optimizing the accuracy of individual agents, they often struggle to generate predictions that are collectively compliant, leading to output trajectories marred by social collisions and environmental violations, thus rendering them impractical for real-world applications. To bridge this gap, we present JACoP: Joint Alignment for Compliant Multi-Agent Prediction, an innovative multi-stage framework that ensures scene-level plausibility. JACoP incorporates an Anchor-Based Agent-Centric Profiler for effective initial compliance filtering and employs a Markov Random Field (MRF) based aligner to formalize the joint selection for scene predictions. By representing inter-agent spatial and social costs as MRF energy potentials, we successfully infer and sample from the joint trajectory distribution, achieving prediction with optimal scene compliance. Comprehensive experiments show that JACoP not only achieves competitive accuracy, but also sets a new standard in reducing both environmental violations and social collisions, thereby confirming its ability to produce collectively feasible and practically applicable trajectory predictions.

1. Introduction

Human Trajectory Prediction (HTP) studies human behavior at a microscopic level by predicting individual movement patterns given historical observations. Due to the complexity of human decision making, along with the flexibility of pedestrian movements, the HTP problem is stochastic by nature, i.e., a given historical observation might be compatible with multiple feasible future outcomes. The rapid development of generative modeling and its successful application in image generation inspired their adoption in the HTP domain.

Previous works[3, 7, 21], leveraging the strong learning capability of SOTA generative models, focused on optimizing prediction accuracy among its K-shot generative outputs. The addition of social and environmental context as input and the dedicated component for modeling interactions further enhance the model's accuracy upper-bound, measured by the displacement of the best prediction among all model outputs (i.e., $min_k ADE/FDE$). To further ensure coverage of a broader range of possible future scenarios and a higher likelihood of recreating ground-truth behavior, previous models also focused on enhancing the diversity of their predictions, either via adoption of alternative learning objective [40] or post-hoc sampling modules [20]. Despite the numerical improvement in the accuracy metric, Our experiment and previsous studies [14, 18] have shown that SOTA models often struggle to generate plausible trajectory predictions free of environmental violation and social collision, rendering the majority generated trajectory predictions unusable as reference for downstream planning tasks. Figure 1 depicts our motivation.

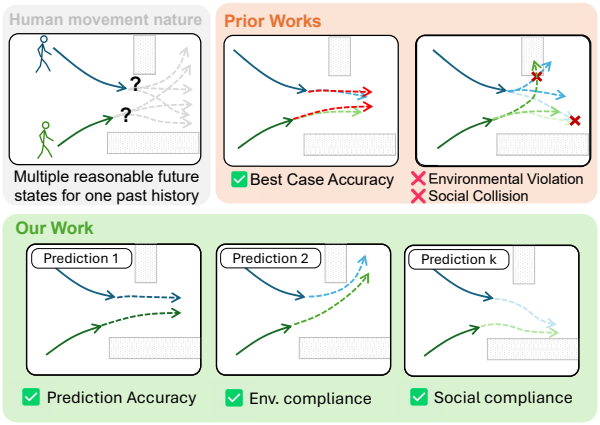


Figure 1. HTP models need environmentally and socially compliant predictions to maximize their utility in downstream tasks. JACoP produce compliant predictions to accomplish this goal.

To address this issue, we propose Joint Alignment for Compliant Prediction (JACoP), a multi-stage framework designed to enforce scene-level compliance across multiple agents simultaneously. The design of JACoP integrates two key elements. First, we introduce an anchor-based agent-

centric profiler, which effectively curates a set of high-quality trajectory prototypes. These prototypes are selected not only to capture a diverse set of agent intentions but also to maintain initial compliance with environmental constraints. This process efficiently constrains the solution space for the subsequent joint sampling stage required for optimal alignment of the scene prediction. Second, we formalize the selection of the final prediction output by inferring a joint trajectory distribution over the selected prototypes through a Markov Random Field (MRF). The MRF explicitly models the spatial and social costs, including joint occurrence likelihood and inter-agent collision penalties, as energy potentials. This allows for the correspondence of the optimal joint trajectory prediction with the lowest-energy configurations of individual prototypes. This mechanism, combined with Gibbs Sampling, effectively models the joint distribution of future trajectories for all scene agents and offers a robust strategy to ensure that the predicted outcomes maintain environmental and social plausibility across the entire scene.

We conduct thorough benchmarking on widely-used trajectory prediction datasets. Additionally, our assessment extends beyond mere accuracy measures of models' best predictions, focusing instead on evaluating their true ability to comprehend scene contexts and simultaneously generate joint scene predictions for all agents. We demonstrate that our approach not only performs competitively in terms of joint accuracy metrics but also sets new benchmarks in reducing environmental violations and social collisions. This confirms that our model effectively closes the gap between precise individual forecasts and the collective feasibility needed for practical human trajectory prediction.

In summary, our contributions are summarized as follows:

- We propose JACoP, a multi-stage trajectory prediction pipeline designed to enforce scene-level compliance across multiple interacting agents simultaneously, bridging the gap between high individual accuracy and collective feasibility.
- We formalize the joint selection of the final predictions by inferring a joint trajectory distribution using an MRF. This model explicitly incorporates spatial and social costs, such as inter-agent collision penalties and joint occurrence likelihood, as energy potentials, guaranteeing low-energy (plausible) configurations.
- We establish new state-of-the-art performance in minimizing crucial metrics for environmental and social compliance. Our work confirms that JACoP successfully generates joint scene predictions that are not only accurate but also consistently practical and usable for downstream planning tasks.

2. Related Works

Human Trajectory Prediction has been a prominent topic in the field of computer vision. Early learning-based methods use recurrent neural networks [1, 10] to model the sequential nature of human trajectory. Later works shift toward using generative models to generate multi-modal predictions better capturing the stochastic nature of human trajectory, using techniques such as Generative Adversarial Network (GAN) [10, 11, 13, 22], Conditional Variational Autoencoders (CVAE) [14, 15, 29, 39], Normalizing flow [5, 26, 27], diffusion model [4, 9, 18, 21] and Flow Matching Model [7].

Several studies developed sampling heuristics to improve prediction diversity. AgentFormer [40] uses a sampling module to enhance diversity from the CVAE module's latent distribution; MemoNet [37] generates a large amount of trajectory samples, performs clustering, and uses cluster centers for maximum diversity, and FEND [35] and AMD [25] focus on sampling long-tail events for better coverage of possible future scenarios.

Recent studies incorporate environmental factors to improve prediction accuracy through scene contexts. Some [14, 20] use rasterized environmental maps to predict waypoints, whereas others [8, 17, 30, 41, 42] use HD maps for vehicle trajectory predictions. Unlike vehicles, humans do not need to strictly follow lane structures or non-drivable areas, therefore lacking direct environmental guidance for accurate predictions. To address this issue, methods including [3, 34] used anchor trajectories as initial proposals, which we also adopted in our model.

Social interaction modeling in the HTP domain is more explored than environmental factors. Research has focused on using social contexts to improve K-shot multi-modal prediction accuracy. Initial studies highlighted social feature extraction, introducing social-pooling to merge information from nearby agents into a target agent's embedding [1, 10]. Later, specialized interaction modules like social graphs, using graph-based learning such as GCN [23, 32] and GAT [11, 13, 42], were developed for multiagent contexts. Other methods [2] furthered this by adding social graph construction as an auxiliary task and using higher-order graphs [6, 12] to better represent group behaviors of pedestrian agents.

Earlier methods use multi-agent contexts for decoding, assuming independent future trajectories for each agent once conditioned on past social contexts. AgentFormer and FJMP [28, 40] address this by inferring from the joint distribution for scene consistency. AgentFormer [40] utilizes autoregressive CVAE to model joint agent motion, while FJMP [28] applies topological sorting and conditional sampling. Our work follows this approach, producing scene-consistent multi-agent predictions by sampling from a joint distribution represented by the MRF. We further enhance

this by optimizing beyond prediction accuracy to ensure the model understands social contexts, achieving both environmental and social compliance.

Previous studies employed $min_k ADE$ to assess model predictions by focusing on the optimal individual agent output. To improve evaluations involving multiple interacting agents, [36] presented JointADE and JointFDE (JADE/JFDE) for predicting pedestrian trajectories. While JADE/JFDE provides a more comprehensive joint performance evaluation, it still only assesses the best sample, lacking a full assessment of trajectory generation quality. Inspired by crowd simulation evaluation metrics, [31] suggested adding qualitative measures like environmental and social collision and diversity metrics to HTP evaluations. These measures assess the overall model capability by considering all multi-modal outputs. In line with this, we incorporate metrics for scene compliance and replicate SOTA models to address this gap in current HTP research.

3. Method

3.1. Overview

Human trajectory prediction (HTP) aims to predict the future movement given a sequence of observed behavior and environmental context. Suppose that there exist a total of N agents in a scene and an observed trajectory sequence $X_i = \{x_i^t | t \in [1, T_o]\}$ for each agent i for T_o steps in 2D global coordinates, where $X = \{X_i | i \in [1, N]\}$. We denote the ground-truth (GT) future trajectory for each agent i as $Y_i = \{y_i^t | t \in [T_o + 1, T_o + T_f]\}$ for T_f steps into the future, and $Y = \{Y_i | i \in [1, N]\}$. Our goal is to jointly predict future trajectories $\hat{Y}$ for all N agents based on historical trajectories X and the environmental context M.

3.2. Agent-Centric Profiler

The Agent-Centric Profiler (ACP) determines the action profile of the target agent through their own historical movement, neighboring agent positions, and environment layout. We use these historical and scene contexts to query against a set of anchor trajectories $\mathbf{Y}^* \subset \mathbb{R}^{M \times T_f \times 2}$ and select K possible future movements as prototype trajectories, producing scores rating for their likelihood in the meantime.

Anchor Trajectory Database We first construct a database of M anchor trajectories from the training set to cover various possible movement types. Inspired by SingularTrajectory [3], we use singular value decomposition (SVD) to compress the GT trajectories Y_i into a lower-dimensional vector $v_i \in \mathbb{R}^{d_s}, d_s = 4$, following the previous work. To do so, we first construct a motion matrix $A \in \mathbb{R}^{N \times 2T_f}$ by flattening and concatenating all the normalized ground-truth future trajectories in the training data. Then we use SVD to decompose the matrix A into

$$A = U \Sigma V \tag{1}$$

where $U \in \mathbb{R}^{N \times d_s}$, $\Sigma \in \mathbb{R}^{d_s \times d_s}$ and $V \in \mathbb{R}^{d_s \times 2T_f}$. We use $V^\top$ to obtain the trajectory compression by

$$v_i = A_i V^\top, \tag{2}$$

where $A_i \in \mathbb{R}^{2T_f}$ is the normalized and flattened version of Y_i for agent i. We then use K-means clustering to group all compressed trajectories into $\mathcal{K}$ clusters and use the cluster centers as the anchor trajectory by transforming them back into the coordinate space to build the database $\mathbf{Y}^* \subset \mathbb{R}^{\mathcal{K} \times T_f \times 2}$.

Feature Extraction Using the anchor trajectory database $\mathbf{Y}^*$, we determine the target agent's motion profile by selecting the K most likely future movements as prototypes using historical context as query. This context includes the agent's past ego status, positions of neighboring agents, and environmental layout. We then project the historical embedding into a future trajectories feature space to choose prototypes whose embeddings are close to this projection.

To obtain the embedding of the historical context $Z_i^{(x)}$ for agent i, we first convert their historical ego status $X_i^{(ego)}$—the temporal sequence of their velocity and heading in polar coordinates—into Fourier features $Z_i^{(ego)} \in \mathbb{R}^{T_o \times D}$ [33], where D is the dimension of embedding space. We then compute the relative spatial-temporal positional embedding $R_{i,j}^{(social)}$ first proposed in the QCNet [42] to encode the historical social context using the position of neighboring agents. $Z_i^{(ego)}$, along with $R_{i,j}^{(social)}$, then is injected into an encoder composed of the factorized attention modules across the temporal and social dimensions, where the ego status embeddings are updated with historical and social contexts.

$$Z_i^{(ego)} = F_\theta(Z_i^{(ego)}, \{R_{i,j}^{(social)}\}_{\forall j \in N_i}), \tag{3}$$

where N_i is the neighbor of agent i. We then takes the embedding corresponding to the last observation step $Z_i^{(ego)}(t) \in \mathbb{R}^D$ to represent an agent's history of past movement and social interactions.

The environmental layout, alongside historical and social contexts, significantly influences pedestrian decision-making. We propose using a distance array $M_i \in \mathbb{R}^{360}$, representing distances to nearby obstacles, to model an agent's local environment. This distance array is encoded into Fourier features and combined with historical data through cross-attention to derive the historical context embedding

$$Z_i^{(x)} = G_\theta(Z_i^{(ego)}(t), M_i). \tag{4}$$

To construct the feature space for future trajectory, we normalize and rotate anchor with respect to the last observed position and heading for each agent. We then obtain the anchor trajectory embedding $Z_\kappa^{(y)} = F_\phi(Y_\kappa^*)$, where $Y_\kappa^* \in \mathbf{Y}^*$ for all $\kappa \in \mathcal{K}$ anchor trajectories using an LSTM

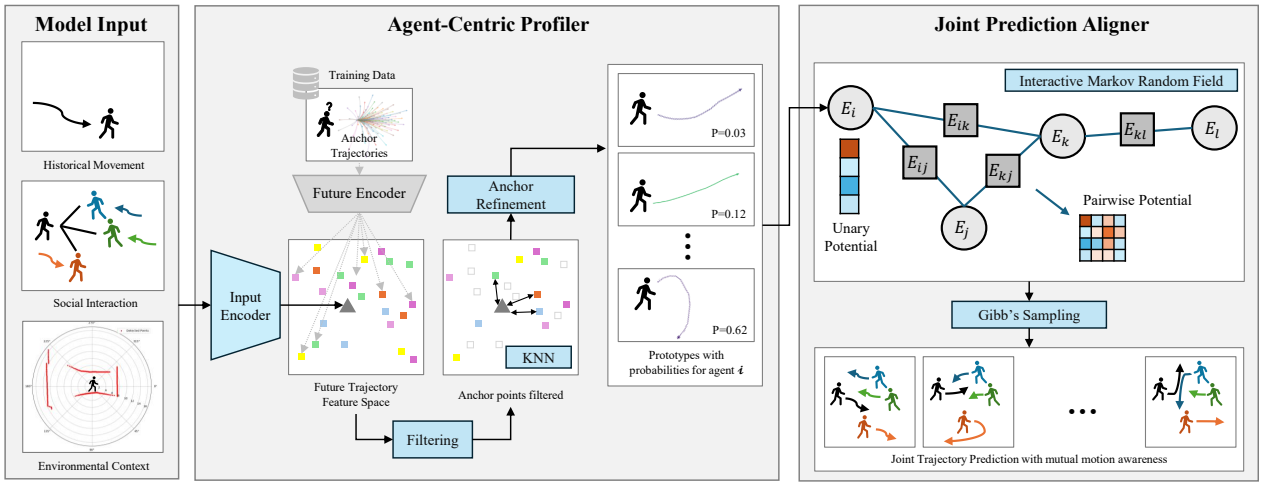


Figure 2. **Model Architecture.** Our framework operates in two stages: (*Left*) Latent embeddings from agents' historical movement, social context, and environment query prototype trajectories, which are filtered and refined. (*Right*) We then use the refined proposals to infer a joint distribution of future trajectories via a Markov Random Field (MRF), with the final scene prediction sampled using Gibbs sampling.

encoder F_ϕ. Both embeddings $Z_i^{(x)}$ and $Z_\kappa^{(y)}$ will capture the motion pattern in their corresponding time horizon. To maintain consistency between these two separate feature spaces (history and future), we further transform the observation embedding $Z_i^{(x)}$ by projecting it into the future trajectory embedding space via an MLP layer h and obtain an updated feature $Z_i^{(x)'} = h(Z_i^{(x)})$.

Prototype Selection We compute the matching score using the cosine similarity values between the updated observation feature and the anchor trajectory embeddings $Z_\kappa^{(y)}$. Specifically, we compute

$$S_i = softmax(\{s_\kappa^i | \kappa = 1, \cdots, \mathcal{K}\}),\ s_\kappa^i = Z_i^{(x)'\top} Z_\kappa^{(y)}, \tag{5}$$

where S_i is the set of matching scores between agent i's observed trajectory against all anchor trajectories. To prevent the selection of anchor trajectories that violate the environmental constraint, we perform zero-out operations to those degenerate ones.[1] We then pick the top-K highest scored prototype as the prototype trajectory set for agent i.

We enhance prediction accuracy by refining prototypes with agent's historical embeddings. For agent i, the customized prototype is derived by adding its observation embedding to the k-th likely prototype $Z_{i_k}^{(y)}$, then decoding it with an LSTM network:

$$\tilde{Y}_{i_k}^* = LSTM(Z_i^{(x)} + Z_{i_k}^{(y)}). \tag{6}$$

This method constructs a customized prototype trajectory set $\tilde{\mathbf{Y}}_i^*$ for agent i, which serves as a candidate set for joint alignment.

[1] Since anchor trajectories are fixed, we can pre-label the ones that violate environmental constraint for target agent i.

3.3. Joint Scene Prediction Alignment

Despite the usage of multi-agent contexts, the prototype selection process still lacks consideration and modeling of the future interaction between agent. Using the highest scoring prototype trajectories of each agent, we can construct an interaction graph $\mathcal{G}$ that estimates the relationship between agents in the future horizon. We then infer the joint distribution of future trajectories $P(Y|X, M)$ over the selected prototypes for all agents in the scene by constructing a Markov Random Field (MRF) from the interaction graph. Inspired by the energy-based formulation of JFP [19], we define the joint trajectory distribution

$$P(Y|X, M) = \frac{1}{\mathcal{Z}} \exp(E(Y|X, M)), \tag{7}$$

where $\mathcal{Z}$ is the normalizing constant. The energy function is defined to be the sum of unary and pair-wise potentials as:

$$E(Y|X, E) = \sum_i E_{unary}(Y_i|X, E) + \sum_{(i,j)\in\mathcal{G}} E_{pairwise}(Y_i, Y_j). \tag{8}$$

For the unary potential $E_i = E_{unary}(Y_i|X, E)$, we directly use the logit of the prototype selection scores s_k^i indicated in Equation (5), where the higher value means that the agent has a higher probability of taking such action. We train a dedicated module to estimate the pairwise potential $E_{ij} = E_{pairwise}(Y_i, Y_j) \in \mathbb{R}^{K\times K}$ for each connected edge between two interacting agents i and j, where

$$E_{ij} = MLP(\tilde{y}_i^*, \tilde{y}_{j@i}^*), \tag{9}$$

and $\tilde{y}_i^*, \tilde{y}_{j@i}^*$ are the embedding of the selected prototype of i and j in the coordinate of the agent i'. We then mask-out

the pair-wise potential value of the colliding pairs of prototypes between agent i and j by assigning large negative values.

Joint Alignment via Gibb's Sampling We apply Gibb's sampling to jointly sample and align prototypes selected from the previous step from the joint distribution estimated by the MRF module. We initialize the sampling process using initial samples from the marginal distribution characterized by the unary potential as:

$$Y^{(0)} \sim P_{unary}, \quad P_{unary} = \frac{1}{\mathcal{Z}} \exp(E_{unary}(Y|X, M)). \tag{10}$$

We then iteratively sample for each agent i. At each sampling step τ, we sample the trajectory of the agent i by conditioning the sample of the other agents' from the previous step as:

$$Y_i^{(\tau)} \sim P\left(Y_i | Y_j^{(\tau-1)}, j \neq i\right). \tag{11}$$

To sample K sets of scene prediction, we save the samples after B steps of the burn-in period. We present the algorithm for the full sampling process in the supplementary material.

3.4. Training Objectives

We train the full pipeline end-to-end using separate loss components for each module. For anchor selection, we apply a focal loss by identifying the prototype with the smallest displacement from the ground-truth trajectory as the GT prototype. The focal loss for the GT prototype's matching score s_m^i is given by:

$$\mathcal{L}_{focal} = -\alpha(1 - s_m^i)^\gamma \log(s_m^i),$$

where α is the balancing parameter and γ controls level of loss contribution from the easier samples.

We train the prototype refinement module using a winner-takes-all strategy. Given the refined prototype trajectories $\tilde{Y}_{i_k}^*$ for $k = 1, \cdots, K$, we define the regression loss that penalizes the displacement error of the most accurate prototype,

$$\mathcal{L}_{regress} = \min_{k=1}^{K} \|\tilde{Y}_{i_k}^* - Y_i\|_2, \tag{12}$$

where Y_i is the GT future trajectory for agent i.

To learn a meaningful feature space for future trajectories, we train the future encoder and decoder by reconstructing GT trajectory Y_i via a reconstruction loss

$$\mathcal{L}_{recon} = \|LSTM(Z_i^{(x)} + F_\phi(Y_i)) - Y_i\|_2. \tag{13}$$

Finally, we make sure that the GT prototype trajectory embedding Z_m^i aligns well with the GT trajectory embedding as

$$\mathcal{L}_{embed} = \|F_\phi(Y_i) - Z_m^i\|_2. \tag{14}$$

Therefore, we define the ACP training loss as a weighted sum,

$$\mathcal{L}_{ACP} = \beta_1 \mathcal{L}_{focal} + \mathcal{L}_{pred} + \mathcal{L}_{recon} + \mathcal{L}_{embed}, \tag{15}$$

where β_1 is a hyperparameter to scale the focal loss. We use $\beta_1 = 100$ for the experiments.

To train the pairwise potential, for each edge in the MRF we mark the prototype pairs between agent i and j with the minimum joint displacement error from the ground-truth as the GT label and then compute the focal loss on the pairwise potential value,

$$\mathcal{L}_{pairwise} = -\alpha(1 - e_{ij}^{(m,n)})^\lambda log(e_{ij}^{(m,n)}), \tag{16}$$

here we assume agent i's mth prototype and agent j's nth prototype shall produce the minimum joint displacement from their GT future. Our final loss function is then composed of both the ACP and pairwise potential loss, $\mathcal{L} = \mathcal{L}_{ACP} + \mathcal{L}_{pairwise}$. Note that, for all focal loss, we use $\alpha = 0.25$ and $\lambda = 2$. We also stop the gradient for $\tilde{y}_i^*, \tilde{y}_{j@i}^*$ in Equation (9) for the pairwise focal loss.

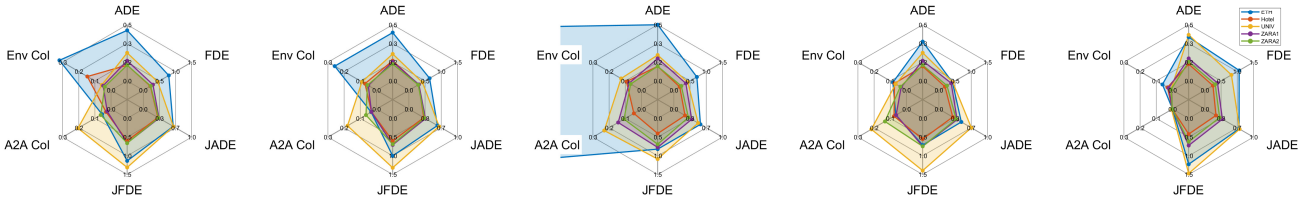
4. Experiments

4.1. Quantitative Evaluation

Dataset We evaluated the performance of our suggested prediction algorithm by performing a benchmark using the widely recognized ETH-UCY datasets [16, 24]. The ETH-UCY dataset contains 1,536 pedestrians from five different scenes: ETH, Hotel, Univ, Zara1, and Zara2; the trajectories are annotated from top-down view footage of surveillance cameras in world coordinates. Here, all results are measured in units of meters. We follow the common training and testing convention first used in Social-GAN [10], with an observation window of 3.2 seconds (8 frames) and a prediction horizon of 4.8 seconds (12 frames) and performed cross-validation against each scene to split the whole dataset into training and testing subsets.

Evaluation Protocols We perform comprehensive evaluations for all benchmarked models based on two main criteria: accuracy and feasibility. To capture the stochastic nature of human movement, we sample k samples from each model and perform the evaluation with $k = 20$. We used Joint ADE/FDE (JADE/JFDE) [36] to measure the model's ability to perform joint prediction across all scene agents. We measure the feasibility of the predicted trajectories by looking at the environmental and agent-to-agent (A2A) collision rate. We also include the commonly used min_kADE/FDE to evaluate the upper bound of each model's predictive capability by measuring the displacement error of the most accurate sample. We provide detailed definition of these metric in the supplementary material.

Analysis We evaluate trajectory prediction using the ETH-UCY dataset and compare our method with Agentformer



(a) AgentFormer (b) GP-Graph (c) EqMotion (d) SingularTrajectory (e) Ours

Figure 3. Radar plot for all evaluation metrics among the five testing splits of ETH-UCY dataset.

Table 1. $\min_k$JADE/JFDE on ETH-UCY dataset with $k = 20$. The best performance is boldfaced and the 2nd place is marked as blue.

Model Avg. # Agents	ETH 2.6	Hotel 3.5	UNIV 25.7	ZARA1 3.7	ZARA 2 6.3
Agentformer	0.619/1.136	0.303/0.603	0.622/1.311	**0.325/0.660**	0.314/0.663
GP-Graph	0.588/1.003	0.307/0.604	0.623/1.319	0.342/0.709	0.322/0.690
EqMotion	0.547/0.822	**0.230/0.405**	**0.499/1.087**	0.367/0.775	0.299/0.692
SingularTrajectory	0.462/0.714	0.286/0.553	0.668/1.395	0.376/0.747	0.356/0.747
Ours	0.704/1.226	0.229/0.420	0.715/1.472	0.361/0.724	**0.304/0.623**

Table 2. Agent-to-Agent Collision Rate (threhold $= 0.2$m) on ETH-UCY dataset. The best performance is boldfaced and the 2nd place is marked as blue.

Model Avg. # Agents	ETH 2.6	Hotel 3.5	UNIV 25.7	ZARA1 3.7	ZARA 2 6.3
Ground Truth	0	0.001	0.035	0	0.007
Agentformer	0.056	0.031	0.205	0.024	0.064
GP-Graph	0.033	0.031	0.179	0.031	0.065
EqMotion	0.52	0.047	0.230	0.144	0.108
SingularTrajectory	0.066	0.077	0.217	0.064	0.133
Ours	**0.00**	**0.001**	**0.006**	**0.001**	**0.001**

Table 3. Environmental Collision Rate on ETH-UCY dataset. The best performance is boldfaced and the 2nd place is marked as blue.

Model	ETH	Hotel	UNIV	ZARA1	ZARA 2
Ground Truth	0	0	0.022	0	0.002
Agentformer	0.32	0.147	0.057	0.049	0.039
GP-Graph	0.258	0.071	0.09	0.051	0.056
EqMotion	0.677	0.093	0.126	0.076	0.067
SingularTrajectory	0.084	0.074	0.067	0.033	0.038
Ours	**0.062**	**0.031**	**0.009**	**0.024**	**0.009**

Table 4. $\min_k$ADE/FDE on ETH-UCY dataset with $k = 20$. The best performance is boldfaced and the 2nd place is marked as blue.

Model	ETH	Hotel	UNIV	ZARA1	ZARA 2
Agentformer	0.45/0.79	0.14/0.22	0.25/0.45	0.18/0.30	0.14/0.23
GP-Graph	0.43/0.64	0.18/0.30	0.24/0.42	**0.17/0.31**	0.15/0.29
EqMotion	0.50/0.72	0.13/0.18	**0.23/0.43**	0.20/0.37	**0.13/0.23**
SingularTrajectory	0.35/0.42	0.13/0.19	0.25/0.43	0.18/0.38	0.14/0.25
Ours	0.59/1.06	0.15/0.25	0.41/0.82	0.20/0.39	0.17/0.32

[40], GP-Graph [2], EqMotion [38], and SingularTrajectory [3]. We assess their JADE/JFDE performance, environmental, and social collision rates by replicating the benchmarked model with the original publication's checkpoints.the

Joint Displacement Error Table 1 presents experiment results of JADE/JFDE on the five partitions of the ETH-UCY dataset. JADE/JFDE evaluates model's ability to simultaneously produce accurate predictions for all scene agents, a more challenging metric compared with the commonly used min_kADE/FDE. The metric requires the model not only to produce accurate individual predictions, but also to align the best samples for all scene agents. Our proposed method demonstrates the ability to match and surpass the overall performance of the SOTA models. Specifically, our model provides a sizable improvement on the Hotel and ZARA2 split.

Social Collision Rate We evaluate the model's understanding of social contexts using the Agent-to-Agent (A2A) Collision Rate, as in Table 2, applying a 0.2m collision threshold to match minimal ground truth collisions. ETH, Hotel, and Zara1 contain scenes with sparse social interactions, each with fewer than five agents per scene, while UNIV and ZARA 2 feature denser scenes. Our model achieves almost collision-free trajectory generation across the whole testing splits, outperforming all benchmarked models. Notably, all SOTA models struggle to generate collision-free predictions, particularly in UNIV and ZARA 2 splits, where up to 20% of the generated trajectories include collisions. This highlights a gap in HTP research, where despite the extensive usage of multi-agent contexts, SOTA models still fail to fully capture socially compliant behaviors. Our method addresses this issue by integrating learned pair-wise potential, collision filtering, and effective sampling strategies to produce trajectories that accurately reflect agents' intentions and maintain social feasibility.

Environmental Violation Table 3 shows the environmental violation rate in the model predictions for the ETH-UCY dataset. The environmental layout represents a significant decision factor in pedestrian movement. In stochastic trajectory prediction, where the model produces multiple prediction samples, all outputs should follow the environmental layout to be viable for downstream tasks. We used the ETH-UCY map annotated by [20] as the gold standard.

Table 5. Ablation study on environmental and social collision rate. The best performance is boldfaced and the 2nd place is marked as blue.

Model			ETH	HOTEL	UNIV	ZARA1	ZARA2	ETH	HOTEL	UNIV	ZARA1	ZARA2
Avg. # Agents			2.6	3.5	25.7	3.7	6.3	2.6	3.5	25.7	3.7	6.3
Env Filter	A2A Filter	Gibbs	Environmental Collision					A2A Collision				
×	×	×	0.154	0.116	0.039	0.077	0.053	0.038	0.054	0.250	0.074	0.130
✓	×	×	0.060	0.037	0.012	0.031	0.020	0.067	0.060	0.253	0.078	0.119
✓	✓	×	**0.058**	0.046	0.013	0.034	0.012	0.066	0.072	0.253	0.082	0.123
✓	✓	✓	0.062	**0.031**	**0.009**	**0.024**	**0.009**	**0.000**	**0.001**	**0.006**	**0.001**	**0.001**

Table 6. Ablation study on JADE/JFDE metric. The best performance is boldfaced and the 2nd place is marked as blue.

Model			ETH	HOTEL	UNIV	ZARA1	ZARA2
Avg. # Agents			2.6	3.5	25.7	3.7	6.3
Env Filter	A2A Filter	Gibbs					
×	×	×	0.703/1.225	0.246/0.446	0.702/1.447	0.397/0.796	0.328/0.672
✓	×	×	0.783/1.305	0.318/0.548	0.626/1.294	0.380/0.766	0.321/0.670
✓	✓	×	0.707/1.234	0.256/0.478	0.680/1.406	0.372/0.744	0.308/0.639
✓	✓	✓	0.704/1.226	**0.229/0.420**	0.715/1.472	**0.361/0.724**	**0.304/0.623**

Both SingularTrajectory and our method incorporate environmental information. Our proposed model achieved minimal environmental collisions using filtering techniques during prototype selection phase, which improved SOTA performance. Compared with SingularTrajectory, which performs environmental correction directly on model predictions, our filtering and refinement approach proved more effective, reducing collisions across the dataset.

Figure 3 shows a radar plot of all evaluation metrics, highlighting our model's significant improvement in feasibility metrics, especially collision avoidance, which is essential for real-world deployment. This is achieved with a minor trade-off in predictive accuracy, reflected by a slight decrease in standard ADE and FDE scores. While this compromise is also evident in $\min_k$ ADE/FDE in Table 4, we emphasize that $\min_k$ ADE/FDE measures only the upper bound of model capability. We strategically prioritize generating a single, highly plausible, and compliant trajectory over optimizing for best-case statistical accuracy, as this better reflects practical application constraints. Further detailed analysis is provided in the supplementary material.

4.2. Qualitative Analysis

We provide qualitative analysis with visualizations of the individual and scene predictions to better showcase the behavior and characteristics of the models in social and environmental interactions.

Figure 4 illustrates two examples of individual predictions from the Hotel and Zara1 scenes. Our model effectively produces diverse, scene-compliant trajectories. In the first example (Figure 4 top), the sample distribution aligns with the environment, predicting straight or rightward paths (in the agent's perspective) as the agent walks alongside a solid barrier on their left. SingularTrajectory uses map gradients for corrections, resulting in compliant but less diverse predictions focusing on forward movement. Agentformer, GP-Graph, and EqMotion do not comply with environmental constraints, mainly predicting leftward turns, which are impractical and unrealistic. In the second example (Figure 4 bottom), our model maintains environmental adherence, generating diverse motions, including 'U-turn' behavior. The other four SOTA methods fail to produce compliant predictions, even for the SingularTrajectory which performs environmental correction; and despite their diverse prediction, they generate infeasible trajectories.

Figure 5 shows two scene predictions from Hotel and Zara1, chosen by the optimal JADE performance among each model's 20 samples. Though SOTA HTP models excel at individual predictions ($min_k ADE$ metric), accurately predicting all agents in a scene remains challenging. Due to the harder nature of joint prediction, model evaluation should allow the generation of alternative future as long as the scene predictions are compliant. In the first example (Figure 5 top), our method mostly aligns with agents' original paths, except for two parallel agents on the right. Our model predicts collision-free paths while keeping their formation, unlike other SOTA models which fail to do so without causing overlaps. In the second example (Figure 5 bottom), our model excels in social comprehension, accurately predicting the intent of all four agents shown in the center of the image, achieving the lowest JADE among benchmarks models, which either miss intentions or predict collisions.

4.3. Ablation Study

Table 5 and Table 6 show the ablation study on environment and social filtering, along with the Gibb's sampling module, regarding collision and joint prediction accuracy metrics. Environmental filtering notably reduces collisions, enhancing model compliance. Without this filtering, the model struggles to maintain environmentally compliant predictions despite using environmental context for prototype refinement.

A2A filtering alone in pair-wise potential is insufficient for socially compliant predictions. We use belief propagation on the MRF to update the unary potential for each agent. We then re-rank the chosen prototype by the updated unary potential values and align the scene prediction using these ranks, (e.g., align all rank 1,2,3...,20 samples). We observe that re-ranking alone does not prevent social collisions or enhance JADE/JFDE performance. Combining A2A filtering with Gibbs sampling significantly reduces A2A col-

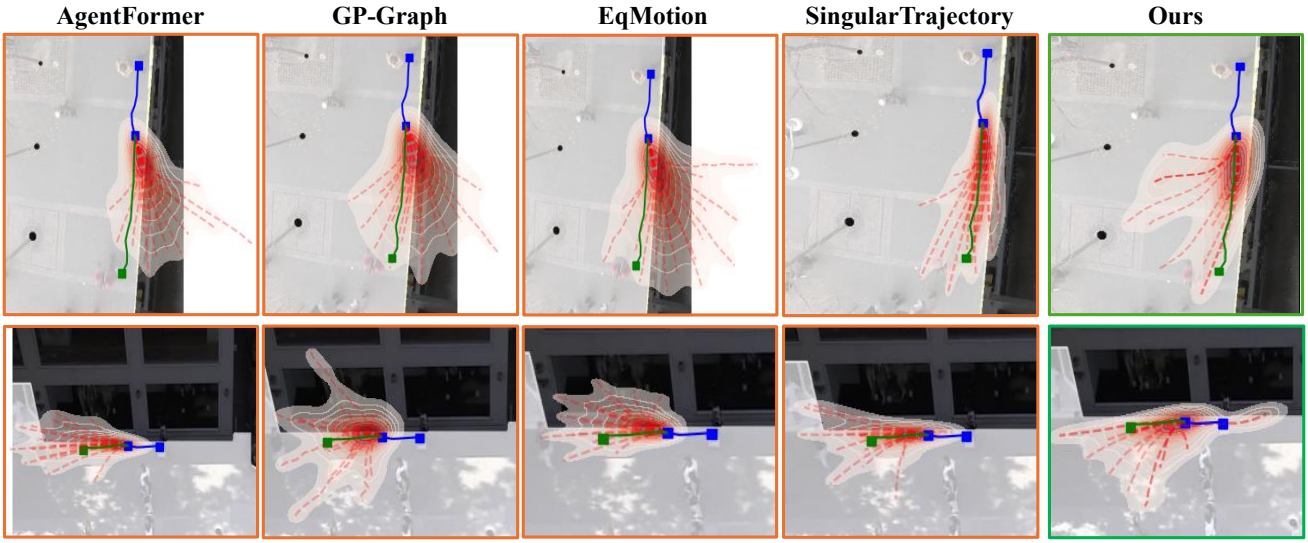


Figure 4. Visualization for all prediction for two individual agents from Hotel (Top) and Zara1 (bottom) splits of ETH-UCY Dataset. Blue line:Historical Trajectory, Green line: Ground Truth Future Trajectory, Red dashed line: Predictions.

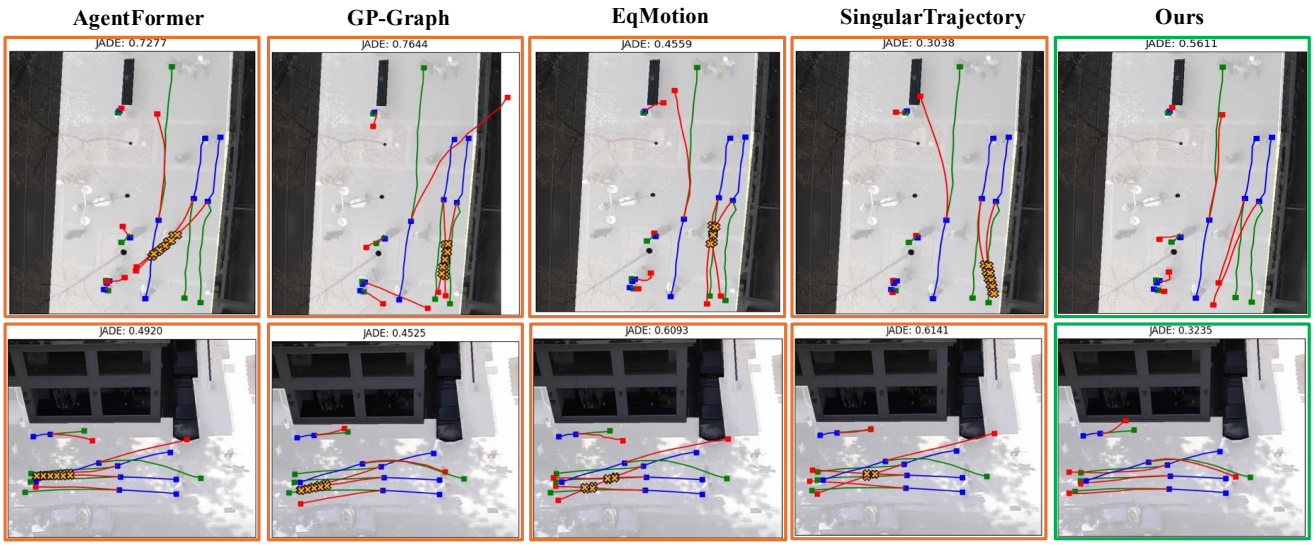


Figure 5. Scene prediction with best JADE performance from Hotel (top) and Zara2 (bottom) split. Blue line:Historical Trajectory, Green line: Ground Truth Future Trajectory, Red dashed line: Prediction, Yellow Cross: Agent-to-Agent Collision.

lisions and improves joint prediction accuracy, showcasing the importance of sampling from a joint trajectory distribution.

5. Conclusion

In this paper, we present JACoP, a novel multi-stage framework that enhances stochastic Human Trajectory Prediction (HTP) by focusing on collective scene-level plausibility. Unlike current top models that often overlook scene consistency, JACoP ensures joint trajectory prediction with minimal breaches of environmental and social constraints. Evaluations show JACoP performs excellently in collective compliance, significantly reducing collisions and boundary violations while maintaining high predictive accuracy. Despite minor declines in accuracy in some data sets and higher time complexity during sampling, JACoP's near-perfect scene compliance makes it suitable for tasks like simulation or crowd behavior analysis, which have less stringent real-time needs.

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. 2

[2] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Learning pedestrian group representations for multi-modal trajectory prediction. In *European Conference on Computer Vision*, pages 270–289. Springer, 2022. 2, 6

[3] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17890–17901, 2024. 1, 2, 3, 6

[4] Guillem Capellera, Antonio Rubio, Luis Ferraz, and Antonio Agudo. Unified uncertainty-aware diffusion for multi-agent trajectory modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22476–22486, 2025. 2

[5] Jiahe Chen, Jinkun Cao, Dahua Lin, Kris Kitani, and Jiangmiao Pang. Mgf: Mixed gaussian flow for diverse trajectory prediction. *Advances in Neural Information Processing Systems*, 37:57539–57563, 2025. 2

[6] Kai Chen, Xiaodong Zhao, Yujie Huang, Guoyu Fang, Xiao Song, Ruiping Wang, and Ziyuan Wang. Socialmoif: Multi-order intention fusion for pedestrian trajectory prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22465–22475, 2025. 2

[7] Yuxiang Fu, Qi Yan, Lele Wang, Ke Li, and Renjie Liao. Moflow: One-step flow matching for human trajectory forecasting via implicit maximum likelihood estimation based distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17282–17293, 2025. 1, 2

[8] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11525–11533, 2020. 2

[9] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17113–17122, 2022. 2

[10] 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. 2, 5

[11] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6272–6281, 2019. 2

[12] Sungjune Kim, Hyung-gun Chi, Hyerin Lim, Karthik Ramani, Jinkyu Kim, and Sangpil Kim. Higher-order relational reasoning for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15251–15260, 2024. 2

[13] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in neural information processing systems*, 32, 2019. 2

[14] 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, 2

[15] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 336–345, 2017. 2

[16] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, pages 655–664. Wiley Online Library, 2007. 5

[17] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 541–556. Springer, 2020. 2

[18] Qingze Tony Liu, Danrui Li, Samuel S Sohn, Sejong Yoon, Mubbasir Kapadia, and Vladimir Pavlovic. Trajdiffuse: A conditional diffusion model for environment-aware trajectory prediction. In *International Conference on Pattern Recognition*, pages 382–397. Springer, 2024. 1, 2

[19] Wenjie Luo, Cheolho Park, Andre Cornman, Benjamin Sapp, and Dragomir Anguelov. JFP: Joint future prediction with interactive multi-agent modeling for autonomous driving. In *Proceedings of The 6th Conference on Robot Learning*, pages 1457–1467. PMLR, 2023. 4

[20] 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, 2, 6

[21] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5517–5526, 2023. 1, 2

[22] Martin Moder and Josef Pauli. Coloss-gan: Collision-free human trajectory generation with a collision loss and gan. In *2021 20th International Conference on Advanced Robotics (ICAR)*, pages 625–632. IEEE, 2021. 2

[23] 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. 2

[24] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*, pages 261–268. IEEE, 2009. 5

[25] Bin Rao, Haicheng Liao, Yanchen Guan, Chengyue Wang, Bonan Wang, Jiaxun Zhang, and Zhenning Li. Amd: Adaptive momentum and decoupled contrastive learning framework for robust long-tail trajectory prediction. *arXiv preprint arXiv:2507.01801*, 2025. 2

[26] Nicholas Rhinehart, Kris M Kitani, and Paul Vernaza. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 772–788, 2018. 2

[27] Nicholas Rhinehart, Rowan McAllister, Kris Kitani, and Sergey Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2821–2830, 2019. 2

[28] Luke Rowe, Martin Ethier, Eli-Henry Dykhne, and Krzysztof Czarnecki. FJMP: Factorized joint multi-agent motion prediction over learned directed acyclic interaction graphs. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13745–13755. IEEE, 2023. 2

[29] 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 16*, pages 683–700. Springer, 2020. 2

[30] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3955–3971, 2024. 2

[31] Samuel S Sohn, Mihee Lee, Seonghyeon Moon, Gang Qiao, Muhammad Usman, Sejong Yoon, Vladimir Pavlovic, and Mubbasir Kapadia. A2x: An agent and environment interaction benchmark for multimodal human trajectory prediction. In *Proceedings of the 14th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–9, 2021. 3

[32] Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 660–669, 2020. 2

[33] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020. 3

[34] B Varadarajan, A Hefny, A Srivastava, KS Refaat, N Nayakanti, A Cornman, K Chen, B Douillard, and CP Lam. D. anguelovet al.,"multipath++: Efficient information fusion and trajectory aggregation for behavior prediction,". *arXiv preprint arXiv*, 2111, 2021. 2

[35] Yuning Wang, Pu Zhang, Lei Bai, and Jianru Xue. Fend: A future enhanced distribution-aware contrastive learning framework for long-tail trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1400–1409, 2023. 2

[36] Erica Weng, Hana Hoshino, Deva Ramanan, and Kris Kitani. Joint metrics matter: A better standard for trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20315–20326, 2023. 3, 5

[37] 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*, pages 6488–6497, 2022. 2

[38] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1410–1420, 2023. 6

[39] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *European Conference on Computer Vision*, pages 511–528. Springer, 2022. 2

[40] 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, 2, 6

[41] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022. 2

[42] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17863–17873, 2023. 2, 3