Diverse Yet Consistent: Context-Guided Diffusion with Energy-Based Joint Refinement for Multi-Agent Motion Prediction

Lei Chu*, and Yuhuan Zhao*
University of Southern California,
Los Angeles, CA, USA
{lc_285, yuhuanzh}@usc.edu

Abstract

Deep generative models have become a promising approach for human motion prediction due to their ability to capture multimodal distributions and represent diverse human behaviors. However, generating predictions that are both diverse and jointly consistent among interacting agents remains challenging. In addition, most existing approaches are primarily evaluated using single-agent (marginal) metrics, which fail to fully reflect the joint dynamics of multi-agent interactions. We propose a diffusion-based framework that improves multi-agent motion prediction by leveraging rich contextual information from historical trajectories. This information is incorporated through a guidance mechanism to enhance the diversity and expressiveness of predicted motions. To further enforce interaction consistency, we introduce an energy-based formulation that refines the joint trajectory distribution while preserving the plausibility of individual trajectories. Extensive experiments on four benchmark datasets demonstrate that our approach consistently outperforms existing methods. Notably, our approach substantially improves both marginal (ADE/FDE) and joint (JADE/JFDE) metrics on ETH/UCY over strong marginal baselines. Compared with prior joint prediction methods, it delivers significant gains in marginal metrics while maintaining competitive joint performance.

1. Introduction

Human trajectory forecasting aims to predict future human movements while accounting for the uncertainty and diversity of possible behaviors [2, 9, 32, 34, 46, 51]. It plays a critical role in applications such as autonomous driving [7, 8, 30, 46, 49], digital health [25, 32], and human–robot interaction [47, 56, 57, 66], where accurately anticipating pedestrian motion is essential for safe decision-making. In multi-agent environments [22, 46], this task requires predicting a distribution of possible future trajectories by modeling both individual motion histories and interactions among agents [67]. Despite significant progress, the problem remains challenging due to the inherent multi modality of human motion, where the same observed past can correspond to multiple plausible future paths across different environments [23].

Classical generative models, such as autoencoder-based approaches, address this challenge by learning compact latent representations of past trajectories and using them to decode and predict future motion. These models efficiently capture motion patterns and can be trained in unsupervised or self-supervised settings, making them easy to extend with techniques such as variational autoencoders [52, 61], attention mechanisms [19, 55], or social interaction modules [2, 35, 61]. However, they often have difficulty capturing multimodal future possibilities, which can lead to averaged trajectories and accumulated errors in long-term predictions. In addition, modeling complex multi-agent interactions often requires extra architectural components. In contrast, diffusion-based generative models [5, 13, 24, 26, 33, 50] naturally model multimodal trajectory distributions by generating multiple plausible futures instead of a single averaged prediction. They can better capture complex motion patterns and interactions, producing diverse and realistic trajectories in crowded or uncertain environments. Nevertheless, diffusion models can exhibit temporal inconsistencies, where stochastic denoising introduces small jitters between time steps, and goal ambiguity, generating trajectories that appear locally plausible but are globally inconsistent with the underlying intent.

To overcome the challenges discussed above, we propose a deep generative modeling approach for human trajectory prediction. The main contributions of this paper are summarized as follows:

- We propose **CODA**, a novel framework for consistent and diverse multi-agent motion prediction. Our approach enhances trajectory generation by incorporating rich contextual information from historical observations and inte-

*Equal contribution, this work was done when the authors were at USC.

grating it into the generative process to improve prediction diversity and expressiveness.
- We introduce a simple yet effective energy-based formulation [11] for joint trajectory refinement, which preserves the plausibility of individual trajectories while improving joint consistency among interacting agents.
- Extensive experiments on four human motion datasets demonstrate that CODA achieves state-of-the-art performance across multiple metrics, effectively balancing the trade-off between marginal metrics (ADE/FDE) and joint metrics (JADE/JFDE) [54], highlighting the importance of modeling diverse yet consistent multi-agent motion.

2. Related Works

2.1. Human Trajectory Prediction

Human trajectory prediction aims to forecast future pedestrian positions from observed motion histories and surrounding interactions. Early approaches relied on physics-based and probabilistic models, such as the Social Force Model[42], Kalman Filters [34], and Hidden Markov Models [12, 38], which describe motion using predefined dynamics but struggle to capture complex behaviors. With the advancement of deep learning, RNN- and LSTM-based models were introduced to learn temporal dependencies in trajectory sequences. To better model social interactions, subsequent work proposed interaction-aware architectures such as Social-LSTM [2] and graph-based methods using Graph Neural Networks (GNNs) [35]. More recently, attention mechanisms and transformer-based models [14, 48] have been explored to capture richer spatial–temporal dependencies and long-range interactions. In parallel, generative approaches, including GANs, VAEs, and diffusion models, have been developed to produce multiple plausible future trajectories, addressing the multi-modality of human motion prediction.

2.2. Interaction Modeling and Enhancement

Existing methods for interaction modeling in human trajectory prediction can be broadly categorized into several groups based on how agent interactions are represented. Rule-based approaches rely on hand-crafted formulations, such as the Social Force Model [42], which describe pedestrian interactions using predefined physical rules [18]. Neighborhood-based aggregation methods capture local interactions by pooling information from nearby agents, as exemplified by social pooling [2]. Pairwise interaction learning methods explicitly model relationships between agent pairs based on their relative positions and motion patterns [21, 63, 68].

More recent work adopts graph-based methods, where Graph Neural Networks (GNNs) represent pedestrians as nodes and their interactions as edges [53], enabling flexible modeling of dynamic crowd structures. Finally, attention-based and transformer-based approaches further enhance interaction modeling by selectively focusing on relevant agents and capturing long-range spatial–temporal dependencies [19, 35, 37]. Building upon these developments, our work introduces richer contextual representations for diffusion-based trajectory generation, enabling diverse yet consistent multi-modal predictions.

2.3. Generation with Guidance

Guided generation in diffusion models is commonly implemented through classifier guidance, which adds gradients from an external classifier trained on noisy samples, and classifier-free guidance (CFG) [16], which combines conditional and unconditional score estimates without requiring a separate classifier. More generally, model- or regressor-guidance methods bias [16, 17, 20] sampling using gradients from learned constraint models. In trajectory prediction, guidance is often realized through goal/intent guidance and scene/map guidance [1, 28, 45], which encourage trajectories to follow likely destinations while respecting environmental constraints [40, 65]. In this work, as illustrated in Fig. 1, we adopt classifier-free guidance and incorporate richer contextual information during generation, while further refining the joint distribution of sampled trajectories with an energy-based model [11, 36, 39], preserving individual trajectory plausibility while improving multi-agent consistency [54].

3. Approach

3.1. Trajectory Prediction Problem

Multi-agent trajectory forecasting aims to predict the future motion of multiple interacting agents based on their historical observations and scene context. Consider a dynamic scene containing N agents observed over T_h time steps. The historical trajectories of all agents are represented as

$$\mathbf{X}_{1:T_h} = \{\mathbf{x}_i^t \mid i = 1, \dots, N,\ t = 1, \dots, T_h\}, \quad (1)$$

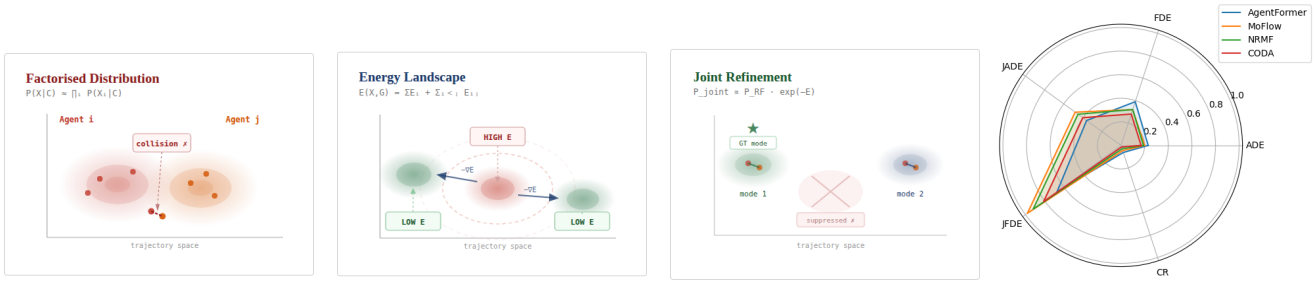
where $\mathbf{x}_i^t \in \mathbb{R}^2$ denotes the 2D spatial position of agent i at time step t. Given these observations, the objective is to predict the future trajectories of all agents over the next T_f time steps,

$$\mathbf{Y}_{T_h+1:T_h+T_f} = \{\mathbf{y}_i^t \mid t = T_h + 1, \dots, T_h + T_f\}, \quad (2)$$

where $\mathbf{y}_i^t \in \mathbb{R}^2$ denotes the ground-truth future position of agent i at time step t. The trajectory forecasting task can therefore be formulated as learning a model $f_\theta(\cdot)$ that estimates the conditional distribution

$$P_\theta(\mathbf{Y}_{T_h+1:T_h+T_f} \mid \mathbf{X}_{1:T_h}, \mathcal{C}), \quad (3)$$

where $\mathcal{C}$ represents additional contextual information such as scene semantics, map priors, and agent interactions.



(a) (b) (c) (d)

Figure 1. Core concept and result of CODA: By incorporating rich interaction context and applying energy-based optimization, CODA improves joint behavior while preserving marginal accuracy, achieving the best performance on marginal metrics (ADE/FDE) and mean Collision Rate (CR), and the second-best results on joint metrics (JADE/JFDE).

Since future trajectories are inherently uncertain and often multi-modal, the model typically predicts a set of K possible future trajectories $\hat{\mathbf{Y}}^{(k)}_{T_h+1:T_h+T_f}, \quad k = 1, \ldots, K$, each corresponding to a plausible future motion hypothesis.

In this section, we introduce the proposed method (Fig. 2), CODA, a diffusion-based framework for joint trajectory modeling. CODA consists of three key modules: (1) Dynamic Context as Guidance Condition (DCGC), which extracts agents' dynamic features as guidance conditions; (2) the Adaptive Condition Integration Module (ACIM), which injects the dynamic context as additional noise to guide embedding generation during diffusion; and (3) Joint Distribution Refinement (JDR), which shifts probability mass toward jointly consistent trajectories while suppressing trajectories that are individually plausible but jointly inconsistent.

3.2. Dynamic Context as Guidance Condition

Recent studies construct guidance conditions from either agent interaction features (non-stationary) [19, 53] or dynamic features (stationary) [1, 12, 18] to guide trajectory generation. However, both approaches have inherent limitations. Non-stationary conditions, represented by dynamic feature embeddings, improve trajectory diversity but often introduce redundant information that may cause deviations from the agent's true intentions. In contrast, stationary conditions maintain intention consistency through agent interest embeddings, but typically rely on a fixed number of representations, which may fail to capture varying intention structures across agents and thus limit trajectory diversity. In this work, inspired by [62, 64], agent-wise features are obtained by applying self-attention [29] to embedded historical trajectories. Additionally, dynamic interaction features are extracted using the context Transformer [62].

3.2.1. Agent-wise Context Extraction

Given the embedded representation of the historical inputs $\mathbf{X}$, we employ a self-attention mechanism [29] to capture dependencies within the sequence. Specifically, a two-layer multilayer perceptron (MLP) is used to compute an attention score matrix that measures the relative importance of different historical observations. The weighted score matrix $\mathcal{W} \in \mathbb{R}^{A\times K}$ is computed as

$$\mathcal{W} = \mathrm{Softmax}\left(\mathrm{MLP}_{4d\times K}\left[\tanh\left(\mathrm{MLP}_{d\times 4d}\left(\mathbf{X}+\mathcal{P}\right)\right)\right]\right), \tag{4}$$

where $\mathcal{P}$ denotes the positional embedding and k represents the number of extracted features, corresponding to the number of generated trajectory hypotheses. The historical information is then aggregated using the weighted score matrix to obtain the interest feature representation:

$$\mathbf{G}_1 \in \mathbb{R}^{k\times N\times d} = \mathcal{W}^{\top} \otimes \mathbf{X}. \tag{5}$$

3.2.2. Global Context Extraction

The global context (dynamic interaction) embedding among agents is extracted using four standard Transformer blocks, similar to the architecture used in [62]. Each block consists of a multi-head self-attention layer followed by a feed-forward network. The resulting global context representation is given by

$$\mathbf{G}_2 \in \mathbb{R}^{1\times N\times d} = \mathrm{Transformer}(\mathbf{X}), \tag{6}$$

where $\mathbf{G}^2$ denotes the learned dynamic interaction embedding and $\mathrm{Transformer}(\cdot)$ represents the stacked Transformer blocks that model dependencies among agents. We repeat $\mathbf{G}^2$ for K times for subsequent use.

Equations Eq. (5) and Eq. (6) describe the extraction of the agent's inherent motion features and mutual interaction features, corresponding to approximations of long-term motion patterns and short-term interaction dynamics. Their combination captures the agent's true motion intention and guides the model to generate realistic trajectory embeddings in the noise space. To preserve the individuality of both features while enriching the guidance signal, we concatenate them to form the final guidance condition:

$$\mathbf{G} \in \mathbb{R}^{K\times A\times 2d} = Concat(\mathbf{G}_1, \mathbf{G}_2, dim = 0), \tag{7}$$

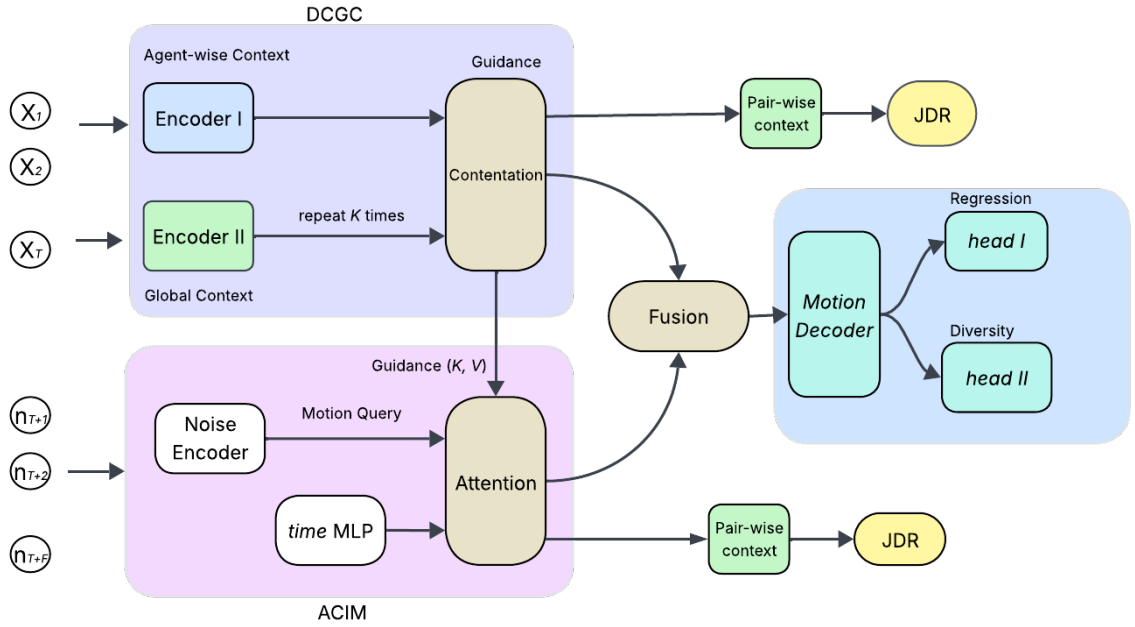


Figure 2. Illustrates the framework of our CODA. It consists of three key modules: (1) Dynamic Context as Guidance Condition (DCGC), which captures agents' dynamic features as guidance conditions; (2) the Adaptive Condition Integration Module (ACIM), which incorporates agent dynamic context from these guidance conditions as additional noise to generate the next embedding during the diffusion generation phase; and (3) Joint Distribution Refinement (JDR), which shifts probability mass toward jointly consistent trajectories while reducing probability assigned to trajectories that are individually plausible but jointly inconsistent.

3.3. Adaptive Context Integration

Existing diffusion-based methods typically employ either MLPs or Transformer encoders for noise (or target) prediction. MLP-based approaches can improve trajectory diversity but lack explicit interaction with guidance conditions, which may introduce bias. In contrast, Transformer-based methods integrate noise representations with agent interaction features through attention weighting, effectively fitting noise in the agent intention space rather than the target space, thereby limiting trajectory diversity.

To address this limitation, we propose an *Adaptive Condition Integration Module* (ACIM) that dynamically injects the guidance condition $\mathbf{G}$ into the noisy target embedding via cross-attention, thereby strengthening conditional guidance during the diffusion process:

$$x_t^G = \text{Softmax}\left(\frac{[(x_t + t_E + A_E)W^Q(GW^K)^\top]}{\sqrt{d}}\right)(GW^V), \tag{8}$$

where t_E and A_E denote the time-step embedding and agent-order embedding, respectively. The guidance-enhanced representation is then fused with the original noisy embedding as

$$x_t = \text{Concat}\left(x_t, x_t^G, \text{dim} = -1\right). \tag{9}$$

3.4. Optimization with Joint Distribution Refinement

3.4.1. Context Guidance based Prediction

Given contextual information $\mathbf{G}$, the goal is to estimate the conditional marginal distribution of the target variable $\mathbf{X}$, denoted as $p(\mathbf{X} \mid \mathbf{G})$. We adopt a conditional diffusion model [16, 40] that learns this distribution through a gradual denoising process. Specifically, a forward diffusion process progressively perturbs the data by adding Gaussian noise:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right), \quad t = 1, \ldots, T, \tag{10}$$

which transforms the data distribution into an isotropic Gaussian. The reverse process learns to recover the clean sample conditioned on context $\mathbf{G}$:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{G}), \tag{11}$$

parameterized by a neural network that predicts the noise component $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{G})$. The model is trained by minimizing the denoising objective

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{G})\|^2\right], \tag{12}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Through iterative denoising conditioned on $\mathbf{G}$, the model learns to sample from the target distribution $p(\mathbf{X} \mid \mathbf{G})$.

To follow the trajectory prediction formulation commonly used in the literature [5, 13, 23, 28], we convert the noise prediction into target prediction and incorporate a best-of-K estimation along with a diversity-encouraging loss:

$$L_{\text{reg}} = \left\| \mathbf{Y} - \hat{\mathbf{Y}}(t, \mathbf{G}) \right\|_2^2 + \min_{k \in \{1,\dots,K\}} \left\| \mathbf{Y} - \widehat{\mathbf{Y}}^{(k)} \right\|_2^2 + \mathcal{L}_{\text{div}}(\hat{\mathbf{Y}}, \mathbf{Y}) \tag{13}$$

In Eq. (13), the first term can be interpreted as a variant of Eq. (12) [5, 13, 23, 28]. The second term corresponds to the well-known best-of- K estimation, while the final term represents the diversity loss. To encourage diversity among the K predicted trajectory modes, we consider the temporal differences between consecutive predictions:$\Delta\hat{y}_i = \hat{y}_{i+1} - \hat{y}_i$ and $\Delta y_i = y_{i+1} - y_i$. The diversity loss is defined as the Kullback-Leibler (KL) divergence between the normalized differences of predicted and ground-truth trajectories:

$$\mathcal{L}_{\text{div}}(\hat{\mathbf{Y}}, \mathbf{Y}) = D_{\text{KL}}(\sigma(\Delta\hat{\mathbf{y}}) \parallel \sigma(\Delta\mathbf{y})) \tag{14}$$

$$= \sum_{i=1}^{T_f - 1} \sigma(\Delta\hat{\mathbf{y}}_i) \log \frac{\sigma(\Delta\hat{\mathbf{y}}_i)}{\sigma(\Delta\mathbf{y}_i)}. \tag{15}$$

The normalization function $\sigma(\cdot)$ is defined using a temperature-scaled softmax:

$$\sigma(\Delta y)_{i,j,k,l} = \frac{\exp(\Delta y_{i,j,k,l}/\tau)}{\sum_{j'=1}^{K} \sum_{k'=1}^{N} \sum_{l'=1}^{T_f} \exp(\Delta y_{i,j',k',l'}/\tau)}, \tag{16}$$

$$\sigma(\Delta \hat{y})_{i,j,k,l} = \frac{\exp(\Delta \hat{y}_{i,j,k,l}/\tau)}{\sum_{j'=1}^{K} \sum_{k'=1}^{N} \sum_{l'=1}^{T_f} \exp(\Delta \hat{y}_{i,j',k',l'}/\tau)}. \tag{17}$$

3.4.2. Joint Distribution Refinement

When the diffusion parameterization factorizes across agents, training yields accurate marginal distributions for each agent but does not explicitly capture their joint dependencies. This limitation motivates us to advocate joint distribution refinement (JDR).

In this work, we build on the observation that diffusion models define a distribution over trajectories and can learn accurate marginals. To address the missing joint dependencies, we introduce an energy-based model (EBM) [11], parameterized by θ, that provides a joint correction to the diffusion distribution. The resulting composed distribution is

$$P_{\text{joint}}(\mathbf{Y} \mid \mathbf{G}) \propto \underbrace{P_{\text{Diff}}(\mathbf{Y} \mid \mathbf{G})}_{\text{good marginals}} \underbrace{\exp(-E_\theta(\mathbf{Y}, \mathbf{G}))}_{\text{joint refinement}}, \tag{18}$$

where $P_{\text{Diff}}(\mathbf{Y} \mid \mathbf{G})$ provides **accurate marginal distributions** via the diffusion model, whereas the energy term $\exp(-E_\theta(\mathbf{Y}, \mathbf{G}))$ acts as an EBM-based correction to better capture **inter-agent dependencies**. Taking the logarithm of Eq. (18) and focusing on the joint refinement part, we obtain

$$\mathcal{L}_{\text{EBM}} = \min_\theta \sum_i E_\theta(\mathbf{Y}_i, \mathbf{G}_i) + \sum_{i<j} E_\theta(\mathbf{Y}_i, \mathbf{Y}_j, \mathbf{G}_{ij}), \tag{19}$$

where the first term evaluates the *individual plausibility* of each agent's prediction by measuring its consistency with the learned marginal distribution. The second term, on the other hand, enforces *joint consistency* among agents by leveraging the **pairwise interaction context**, $\mathbf{G}_{ij}$, which encodes relational dependencies between agents. This interaction-aware mechanism encourages the predicted trajectories to remain not only individually realistic but also mutually compatible within the multi-agent environment [11, 36, 39]. We note that the refinement in Eq. (18) is related to prior work [10], where diffusion models are improved using score-based gradients. In contrast, our method performs refinement at the distribution level rather than the score level.

With the analysis above, we obtain the training loss used for our network. The overall training objective can be expressed as:

$$\mathcal{L} = \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{EBM}}\mathcal{L}_{\text{EBM}} \tag{20}$$

3.4.3. Model Optimization and Discussion

In our approach, trajectories are first normalized using min–max scaling to map future relative motions to the range $[-1, 1]$, which helps stabilize training dynamics. For most datasets, we adopt the Social Transformer as the backbone architecture, while an MLP backbone is used for the NBA dataset. The transformer-based encoders incorporate skip connections and share a common configuration with 128 hidden features, a feed-forward dimension of 512, eight attention heads, and four stacked layers. The diffusion model generates samples through a 100-step denoising ODE process. For the flow time scheduler, we employ a logit-normal distribution. All experiments are conducted on an NVIDIA GeForce RTX 5090 GPU using the AdamW optimizer in PyTorch, with a weight decay of 0.01.

From an efficiency perspective, the proposed formulation is compatible with existing one-step and few-step diffusion models. Empirically, we find that a student model implemented as a one-step diffusion model achieves performance comparable to that of the teacher model when trained using Maximum Likelihood Estimation (MLE)-based distillation objectives, including KL divergence, Maximum Mean Discrepancy (MMD), and Chamfer distance. As optimization efficiency is not the primary focus of this work, we omit additional case studies of the one-step variant.

4. Experiments

4.1. Datasets

We evaluate the proposed method on four widely used trajectory prediction benchmarks: ETH/UCY, SDD, NBA, and JRDB. The ETH/UCY dataset contains five subsets (ETH, HOTEL, UNIV, ZARA1, ZARA2); following the standard leave-one-out protocol, we train on four subsets and test on the remaining one, predicting 12 future frames (4.8 s) from 8 observed frames (3.2 s). The SDD dataset [41] provides bird's-eye-view pedestrian trajectories in pixel coordinates without projection matrices; we predict 12 future frames from 8 observations and report results in both pixel and metric units. The NBA dataset [6] contains trajectories of 10 players and the ball captured by the SportVU system; we predict 20 future frames (4.0 s) conditioned on 10 observed frames (2.0 s), where frequent abrupt intention changes make trajectories more complex than typical pedestrian scenarios. The JRDB dataset [44] is a large-scale egocentric benchmark collected by a mobile social robot across diverse environments; we use the Social-Transmotion split [43] for deterministic evaluation and the official challenge splits for stochastic prediction. As trajectories are annotated in camera coordinates while the robot is moving, we convert them to a global frame using odometry from rosbags. Following the official protocol, the model predicts 12 future frames from 9 observations at 2.5 Hz.

4.2. Baselines

We compare the proposed CODA model with several state-of-the-art approaches, including NPSN [3], S-GAN[15], GroupNet [58], LED [33], TUTR [48], EqMotion [60], EigenTraj [4], SingularTraj [5], Evo-Graph [35], Y-net [32], PECNet [31], SocialVAE [61], MemoNet [59], LRR [27], MOFLOW [13] and NMFT [12], on datasets where comparable results are publicly available. For the most recent methods, MOFLOW [13] and NMFT [12], we reproduce the reported results using their official codebases to ensure a fair comparison. The performance of the remaining methods is taken directly from their respective publications. Note that the set of compared methods may vary across datasets depending on the availability of reported results.

4.2.1. Evaluation protocols

We employ both marginal (ADE and FDE) and joint (JADE and JFDE) metrics for evaluation [54], which differ in the order of aggregation over samples and agents. The marginal metrics compute the minimum over K predicted trajectories of the average displacement across time steps and the final-step displacement, respectively, on a per-agent basis. In contrast, joint metrics first average displacement errors across all agents within each predicted sample and then select the minimum over K samples. This reordering, though subtle, is critical, as it enforces sample-level consistency and prevents combining predictions of different agents from different samples during evaluation.

Table 1. Minimum ADE/FDE comparison across datasets with K=20 samples. Lower values indicate better performance. The best results are highlighted in bold, and the second-best results are underlined.

Methods	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg.
S-GAN [15]	0.88/1.66	0.46/0.92	0.64/1.34	0.38/0.82	0.29/0.60	0.53/1.07
Trajectron++ [46]	0.67/1.18	0.19/0.28	0.30/0.54	0.25/0.41	0.18/0.32	0.32/0.55
PECNet [31]	0.56/0.99	0.19/0.33	0.34/0.63	0.24/0.47	0.18/0.35	0.30/0.55
Y-Net [32]	0.40 / 0.57	0.12/0.19	0.31/0.60	0.26/0.49	0.20/0.39	0.26/0.45
MemoNet [59]	0.41/0.64	0.11/0.17	0.24/ 0.43	0.18/0.32	0.14/0.25	0.22/0.36
View Vertically [55]	0.57/0.69	0.12/0.19	0.29/0.50	0.20/0.36	0.15/0.26	0.27/0.40
Joint View Vertically [54]	0.70/0.79	0.13/0.20	0.27/0.47	0.22/0.36	0.14/0.25	0.29/0.41
AgentFormer [64]	0.45/0.75	0.14/0.23	0.25/0.45	0.18 / 0.30	0.14/0.24	0.23/0.39
Joint AgentFormer [54]	0.47/0.79	0.14/0.21	0.29/0.51	0.19/0.32	0.14/0.24	0.25/0.41
LRR [27]	N/A	N/A	N/A	N/A	N/A	N/A
MoFlow [13]	0.40/0.57	0.11/0.17	0.23/**0.39**	**0.15**/0.26	**0.12/0.22**	0.20/0.32
NRMF [12]	0.26/**0.37**	0.11/0.17	0.28/0.49	0.18/0.30	0.14/0.25	0.19/0.32
CODA (Ours)	**0.24/0.37**	**0.10/0.15**	**0.22/0.39**	**0.15/0.25**	**0.12/0.22**	**0.17/0.28**

4.3. Quantitative Results

ETH/UCY: Tab. 1 reports the minimum ADE/FDE on ETH/UCY with $K = 20$ samples. CODA achieves the best overall performance, obtaining the lowest average error (0.17/0.28), corresponding to a 15.0% / 12.5% improvement over MoFlow (0.20/0.32) and a 10.5% / 12.5% improvement over MRF (0.19/0.32). Our method achieves the best ADE on ETH (0.24) and UNIV (0.22), and the lowest or tied-lowest FDE on ZARA1 (0.25) and ZARA2 (0.22). Compared with earlier generative models such as S-GAN, PECNet, and MemoNet, the improvements are substantially larger (over 30–60% on average), demonstrating significantly enhanced trajectory accuracy and long-term prediction fidelity. The consistent gains across scenes indicate stronger multimodal modeling and more effective sample selection under the marginal evaluation protocol.

Tab. 2 presents joint evaluation results, where the best sample is selected after averaging errors across all agents, imposing stricter multi-agent consistency. Under this setting, CODA remains competitive, achieving the best performance on UNIV (0.52 JADE) and strong results across other subsets, with a competitive overall average (0.40/0.81). While some methods (e.g., AgentFormer variants and Joint VV) perform well on specific scenes, our approach maintains stable performance across environments. Moreover, the relatively small gap between marginal and joint metrics suggests that CODA produces coherent multi-agent predictions within each sampled trajectory, rather than relying on per-agent best-case selection. Overall, CODA generates diverse and accurate trajectories while preserving strong inter-agent consistency, demonstrating robustness across evaluation protocols.

NBA: Tab. 3 presents the temporal evaluation on the NBA dataset under both marginal (ADE/FDE) and joint (JADE/JFDE) metrics. Across all prediction horizons

Table 2. Quantitative comparison using min $JADE_{20}/JFDE_{20} \downarrow$ with $K = 20$ samples)

Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg.
S-GAN [15]	0.92/1.7	0.48/0.95	0.74/1.57	0.44/1.0	0.36/0.79	0.59/1.21
Trajectron++ [46]	0.73/1.3	0.24/0.42	0.61/1.32	0.36/0.71	0.29/0.63	0.45/0.87
PECNet [31]	0.62/1.1	0.29/0.59	0.67/1.42	0.41/0.90	0.37/0.84	0.47/0.97
Y-Net [32]	0.50/0.78	0.21/0.39	0.70/1.56	0.49/1.04	0.49/1.10	0.48/0.97
MemoNet [59]	0.50/0.86	0.22/0.42	0.69/1.47	0.35/0.72	0.39/0.86	0.43/0.87
View Vertically [55]	0.56/**0.78**	0.20/0.33	0.65/1.31	0.33/0.65	0.30/0.60	0.41/0.73
Joint View Vertically [54]	0.65/0.84	**0.19/0.31**	**0.52/1.09**	0.33/0.63	0.27/0.55	0.39/0.68
AgentFormer [54]	**0.48**/0.79	0.24/0.46	0.62/1.31	0.29/0.56	0.30/0.62	0.38/0.75
Joint AgentFormer [54]	0.49/0.80	0.19/0.32	0.59/1.22	**0.27/0.51**	**0.25/0.51**	**0.36/0.67**
LRR [27]	0.51/0.91	0.19/0.33	0.61/1.25	0.33/0.65	0.28/0.59	0.39/0.74
MoFlow [13]	0.71/1.36	0.34/0.66	0.59/1.12	0.41/0.93	0.38/0.87	0.48/0.98
NRMF [12]	0.57/1.11	0.27/0.50	0.64/1.31	0.43/0.94	0.34/0.74	0.45/0.92
CODA (Ours)	0.56/1.07	0.23/0.42	**0.52**/1.10	0.39/0.82	0.30/0.66	0.40/0.81

Table 3. Comparison with state-of-the-art methods on the NBA dataset.

	Time	GroupNet	MID	LED	MOFLOW	NRMF	CODA
ADE	1.0s	0.26/0.34	0.28/0.37	0.21/0.28	0.18/0.25	**0.16/0.24**	0.17/**0.24**
/FDE	2.0s	0.49/0.70	0.51/0.72	0.44/0.64	0.34/0.48	0.34/0.50	**0.33/0.45**
	3.0s	0.73/1.02	0.71/0.98	0.69/0.95	0.51/0.68	0.53/0.75	**0.50/0.66**
	4.0s	0.96/1.30	0.96/1.27	0.94/1.21	0.70/0.89	0.75/0.97	**0.69/0.87**
JADE	1.0s	N/A	N/A	N/A	0.37/0.68	**0.33/0.61**	0.36/0.67
/JFDE	2.0s	N/A	N/A	N/A	0.81/1.62	**0.73/1.46**	0.79/1.59
	3.0s	N/A	N/A	N/A	1.26/2.50	**1.15/2.21**	1.23/2.47
	4.0s	N/A	N/A	N/A	1.69/3.31	**1.53/2.79**	1.67/3.28

(1.0s–4.0s), CODA demonstrates consistently strong performance. Under marginal metrics, CODA achieves the best results at longer horizons, obtaining the lowest ADE/FDE at 3.0s (0.50/0.66) and 4.0s (0.69/0.87), while remaining competitive at shorter horizons. Although NRMF attains the best performance at 1.0s (0.16/0.24), CODA maintains comparable accuracy and outperforms the other baselines as the prediction horizon increases. Overall, these ADE/FDE results indicate that CODA provides the most robust marginal trajectory prediction, particularly for longer-term forecasting.

For the joint metrics, NRMF achieves the best results across all prediction horizons. CODA ranks second in most cases. The widening gap at longer horizons indicates that NRMF remains a strong competitor to CODA, and that both methods are effective at modeling long-term dynamics and complex inter-agent interactions in NBA trajectories. Overall, the JADE/JFDE results show that both methods model multi-agent dependencies well, with NRMF having a slight advantage in joint forecasting accuracy.

SDD: Tab. 4 reports the quantitative comparison on the SDD dataset under both marginal (ADE/FDE) and joint (JADE/JFDE) metrics. CODA achieves the best marginal performance, obtaining the lowest ADE (6.87) and FDE (10.86), outperforming strong recent baselines such as MRF (7.10/11.11), MoFlow (7.50/11.96), and ET+HighGraph (7.81/11.09). These improvements are notable given the highly dynamic nature of NBA trajectories, which involve abrupt motion changes and complex player interactions.

Under joint metrics, AgentFormer achieves the best JADE (9.56) and Y-Net the lowest JFDE (16.01), indicating stronger optimization for joint trajectory coherence. In contrast, recent generative approaches—including MoFlow, NRMF, and CODA—primarily optimize marginal trajectory accuracy, which can reduce joint consistency. Overall, CODA establishes state-of-the-art marginal performance on NBA while maintaining competitive joint results, demonstrating its effectiveness in modeling complex, interaction-rich dynamics.

Table 4. Comparison with state-of-the-art methods on the SDD dataset.

Methods	Venue	ADE	FDE	JADE	JFDE
S-GAN	CVPR'18	12.74	22.65	13.76	24.84
Trajectron++	ECCV'20	10.18	15.76	11.36	18.21
PECNet	ECCV'20	9.34	16.10	10.82	19.48
Y-Net	ICCV'21	8.15	12.80	9.67	**16.01**
MemoNet	CVPR'22	7.97	12.82	9.59	16.43
View Vertically	ECCV'22	9.34	14.67	10.75	17.45
Joint View Vertically	CVPR'23	9.62	15.07	10.92	17.70
AgentFormer	CVPR'21	8.01	13.24	9.67	16.92
Joint AgentFormer	CVPR'23	8.25	13.74	**9.56**	16.59
TUTR	ICCV'23	7.76	12.69	-	-
ET+HighGraph	CVPR'24	7.81	11.09	-	-
MoFlow	CVPR'25	7.50	11.96	18.97	37.46
NRMF	ICLR'25	7.10	11.11	16.18	32.55
CODA (Ours)	TBD	**6.87**	**10.86**	13.96	26.31

JRDB: Tab. 5 reports time-horizon evaluation on JRDB using both marginal (ADE/FDE) and joint (JADE/JFDE) metrics. Across prediction intervals from 1.2s to 4.8s, CODA consistently achieves competitive or superior performance. Under marginal metrics, it attains the best results at 1.2s (0.04/0.05) and remains tied for the lowest errors at longer horizons, reaching 0.11/0.17 at 3.6s and 0.15/0.23 at 4.8s, indicating strong short-term precision and stable long-term forecasting. Under joint metrics, CODA demonstrates improved multi-agent consistency over MOFLOW and NRMF, particularly at longer horizons; at 4.8s, it achieves the lowest JADE/JFDE (0.34/0.64), outperforming MOFLOW (0.35/0.67) and NRMF (0.37/0.69). The widening gap at extended horizons suggests better preservation of global scene coherence. Overall, CODA maintains strong accuracy across horizons while achieving superior long-term joint consistency, highlighting its effectiveness in modeling dynamic, egocentric multi-agent environments.

Table 5. Comparison with state-of-the-art methods on the JRDB dataset.

	Time	LED	MOFLOW	NRMF	CODA
ADE/FDE	1.2s	0.05/0.07	0.04/0.06	**0.04/0.05**	**0.04/0.05**
	2.4s	0.09/0.14	**0.07/0.11**	0.08/**0.11**	**0.07/0.11**
	3.6s	0.14/0.21	**0.11/0.17**	**0.11/0.17**	**0.11/0.17**
	4.8s (Total)	0.18/0.28	**0.15/0.23**	**0.15/0.23**	**0.15/0.23**
JADE/JFDE	1.2s	N/A	0.10/0.15	**0.09**/0.16	**0.09/0.14**
	2.4s	N/A	0.18/0.32	0.19/0.35	**0.17/0.31**
	3.6s	N/A	0.27/0.50	0.28/0.52	**0.25/0.47**
	4.8s (Total)	N/A	0.35/0.67	0.37/0.69	**0.34/0.64**

4.4. Qualitative Results

This section presents a qualitative comparison of multi-modal trajectory predictions for Scene 2, Agent 8 from the NBA dataset. Each column corresponds to a different method (MOFLOW, NRMF, and CODA), visualizing the observed trajectories, ground-truth futures, sampled predictions, and the mean predicted trajectory. The reported standard deviation reflects the dispersion of sampled trajectories and provides an indication of predictive diversity.

As shown in Fig. 3, MOFLOW generates diverse trajectories that largely align with the ground truth but exhibit mild dispersion in highly dynamic cases. NRMF produces more concentrated predictions with controlled variance while maintaining plausible future directions. CODA achieves similar diversity with better alignment to the ground truth and interaction context, resulting in improved structural coherence across agents. Overall, while all methods capture complex interaction-driven sports dynamics, CODA demonstrates stronger joint behavioral consistency.

Similar observations can be made in Fig. 4. NRMF produces diverse trajectories but occasionally deviates from the dominant motion pattern, resulting in larger displacement errors. MOFLOW more closely follows the ground truth while preserving trajectory diversity. CODA strikes a balance between diversity and structural coherence, capturing the overall motion trend while maintaining interaction-consistent variations. Although its ADE (0.2371) is slightly higher than that of MOFLOW in this example, CODA achieves a lower JADE (0.9299), indicating improved joint behavioral consistency across agents.

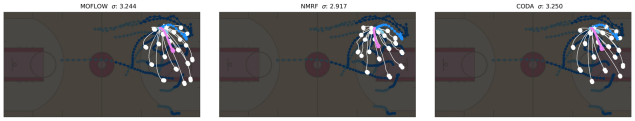


Figure 3. Prediction samples (T_h=20) on the NBA dataset. Light blue shows historical trajectories, dark blue shows future ground truth, white curves are sampled predictions, and the violet curve denotes the mean estimate.

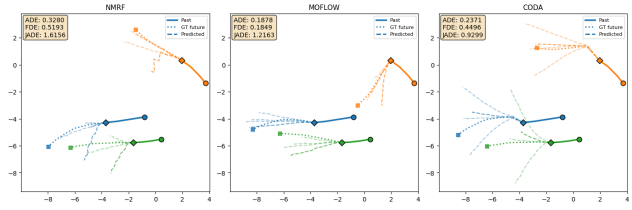


Figure 4. Qualitative trajectory prediction results on the Univ dataset.

Table 6. Ablation study evaluating the components of the proposed method.

DCGC	ACIM	JDR	ADE/FDE	JADE/JFDE
✓	×	×	0.740/0.923	1.809/3.485
✓	✓	×	0.724/0.901	1.764/3.421
×	×	✓	0.716/0.897	1.665/3.273
✓	✓	✓	0.694/0.873	1.671/3.283

4.5. Ablation Study

We conduct an ablation study on the NBA dataset to examine the contributions of the key components in our framework—DCGC, ACIM, and JDR—using both marginal metrics (ADE/FDE) and joint metrics (JADE/JFDE). As reported in Tab. 6, removing DCGC or ACIM leads to clear increases in ADE and FDE, indicating that these modules play an essential role in improving agent-wise trajectory prediction by capturing richer contextual and interaction information. In contrast, removing JDR results in only minor changes in ADE/FDE but causes noticeable degradation in JADE/JFDE. This result suggests that JDR mainly contributes to modeling joint dynamics and coordination among agents. Overall, the ablation results demonstrate a clear functional distinction: DCGC and ACIM primarily enhance marginal prediction accuracy, whereas JDR is crucial for improving joint trajectory consistency.

5. Conclusion

In this work, we propose CODA, a framework for improving multi-agent motion prediction by incorporating rich contextual information from historical trajectories into the generative process, enabling more diverse and expressive predictions. We further introduce a simple yet effective formulation that refines the joint trajectory distribution using an energy-based model, preserving the plausibility of individual trajectories while enhancing joint consistency among interacting agents. Extensive experiments on four benchmark datasets demonstrate consistent improvements over state-of-the-art approaches. Overall, these results highlight the effectiveness of CODA in jointly modeling diversity and consistency for multi-agent motion prediction.

Limitations. We leverage different levels of data-driven context, achieving SOTA performance on ADE/FDE while attaining competitive results on the joint metrics JADE, JFDE, and CR. However, the additional effort devoted to robust context modeling increases the training time. All methods still have substantial room for improvement on the CR metric, although this issue could be significantly alleviated by incorporating LiDAR data. Furthermore, the current prediction pipeline relies solely on coordinate inputs; integrating richer physical context—such as traversable areas—could further enhance performance.

References

[1] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *ICCV*, pages 13390–13400, 2021. 2, 3

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, pages 961–971, 2016. 1, 2

[3] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Non-probability sampling network for stochastic human trajectory prediction. In *CVPR*, pages 6477–6487, 2022. 6

[4] Inhwan Bae, Jean Oh, and Hae-Gon Jeon. Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *ICCV*, pages 10017–10029, 2023. 6

[5] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *CVPR*, pages 17890–17901, 2024. 1, 5, 6

[6] Dan Cervone, Alexander D'amour, Luke Bornn, and Kirk Goldsberry. Pointwise: Predicting points and valuing decisions in real time with nba optical tracking data. In *Proceedings of the 8th MIT Sloan Sports Analytics Conference, Boston, MA, USA*, 2014. 6

[7] Lei Chu, Daoud Burghal, Rui Wang, Michael Neuman, and Andreas F Molisch. Context-conditioned spatio-temporal predictive learning for reliable v2v channel prediction. *TITS*, 2026. 1

[8] Alexander Cui, Sergio Casas, Abolfazl Sadat, Renjie Liao, and Raquel Urtasun. Lookout: Diverse multi-future prediction and planning for self-driving. In *CVPR*, pages 16107–16116, 2021. 1

[9] Nachiket Deo and Mohan M. Trivedi. Convolutional social pooling for vehicle trajectory prediction. In *CVPR*, pages 1468–1476, 2018. 1

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 5

[11] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *NeurIPS*, 32, 2019. 2, 5

[12] Zilin Fang, David Hsu, Gim Hee Lee, and Gim Hee Lee. Neuralized markov random field for interaction-aware stochastic human trajectory prediction. In *ICLR*, 2025. 2, 3, 6, 7

[13] 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 *CVPR*, pages 17282–17293, 2025. 1, 5, 6, 7

[14] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. In *ICPR*, pages 10335–10342. IEEE, 2021. 2

[15] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, pages 2255–2264, 2018. 6, 7

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *ICLR*, 2022. 2, 4

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[18] De-An Huang and Kris M Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *ECCV*, pages 489–504. Springer, 2014. 2, 3

[19] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *CVPR*, pages 6272–6281, 2019. 1, 2, 3

[20] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *CVPR*, pages 12469–12478, 2024. 2

[21] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022. 2

[22] Jaewoo Jeong, Daehee Park, and Kuk-Jin Yoon. Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning. In *CVPR*, pages 1617–1628, 2024. 1

[23] Jaewoo Jeong, Seohee Lee, Daehee Park, Giwon Lee, and Kuk-Jin Yoon. Multi-modal knowledge distillation-based human trajectory forecasting. In *CVPR*, pages 24222–24233, 2025. 1, 5

[24] Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *CVPR*, pages 9644–9653, 2023. 1

[25] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *TITS*, 23(7):7386–7400, 2021. 1

[26] Rongqing Li, Changsheng Li, Dongchun Ren, Guangyi Chen, Ye Yuan, and Guoren Wang. Bcdiff: Bidirectional consistent diffusion for instantaneous trajectory prediction. *NeurIPS*, 36:14400–14413, 2023. 1

[27] Haotian Lin, Yixiao Wang, Mingxiao Huo, Chensheng Peng, Zhiyuan Liu, and Masayoshi Tomizuka. Joint pedestrian trajectory prediction through posterior sampling. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5672–5679, 2024. 6, 7

[28] Xiaotong Lin, Tianming Liang, Jianhuang Lai, and Jian-Fang Hu. Progressive pretext task learning for human trajectory prediction. In *ECCV*, pages 197–214. Springer, 2024. 2, 5

[29] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *ICLR*, 2017. 3

[30] Dongrui Liu, Chuanchaun Chen, Changqing Xu, Robert C Qiu, and Lei Chu. Self-supervised point cloud registration with deep versatile descriptors for intelligent driving. *TITS*, 24(9):9767–9779, 2023. 1

[31] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint

conditioned trajectory prediction. In *ECCV*, pages 759–776, 2020. 6, 7

[32] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *CVPR*, pages 15233–15242, 2021. 1, 6, 7

[33] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *CVPR*, pages 5517–5526, 2023. 1, 6

[34] Mancheng Meng, Ziyan Wu, Terrence Chen, Xiran Cai, Xiang Zhou, Fan Yang, and Dinggang Shen. Forecasting human trajectory from scene history. In *NeurIPS*, pages 24920–24933, 2022. 1, 2

[35] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *CVPR*, pages 14424–14432, 2020. 1, 2, 6

[36] Bo Pang, Tianyang Zhao, Xu Xie, and Ying Nian Wu. Trajectory prediction with latent belief energy-based model. In *CVPR*, pages 11814–11824, 2021. 2, 5

[37] Yusheng Peng, Gaofeng Zhang, Xiangyu Li, and Liping Zheng. Stirnet: A spatial-temporal interaction-aware recursive network for human trajectory prediction. In *CVPR*, pages 2285–2293, 2021. 2

[38] Shaojie Qiao, Dayong Shen, Xiaoteng Wang, Nan Han, and William Zhu. A self-adaptive parameter selection trajectory prediction approach via hidden markov models. *TITS*, 16(1): 284–296, 2014. 2

[39] Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann Cun. Efficient learning of sparse representations with an energy-based model. *NeurIPS*, 19, 2006. 2, 5

[40] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *CVPR*, pages 13756–13766, 2023. 2, 4

[41] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, pages 549–565. Springer, 2016. 6

[42] Andrey Rudenko, Luigi Palmieri, and Kai O Arras. Joint long-term prediction of human motion using a planning-based social force approach. In *ICLR*, pages 4571–4577. IEEE, 2018. 2

[43] Saeed Saadatnejad, Yang Gao, Kaouther Messaoud, and Alexandre Alahi. Social-transmotion: Promptable human trajectory prediction. In *ICLR*, 2023. 6

[44] Saeed Saadatnejad, Yang Gao, Hamid Rezatofighi, and Alexandre Alahi. Jrdb-traj: A dataset and benchmark for trajectory forecasting in crowds. *arXiv preprint arXiv:2311.02736*, 2023. 6

[45] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2

[46] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *ECCV*, pages 683–700. Springer, 2020. 1, 6, 7

[47] Alessio Sampieri, Guido Maria D'Amely di Melendugno, Andrea Avogaro, Federico Cunico, Francesco Setti, Geri Skenderi, Marco Cristani, and Fabio Galasso. Pose forecasting in industrial human-robot collaboration. In *ECCV*, pages 51–69. Springer, 2022. 1

[48] Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *ICCV*, pages 9675–9684, 2023. 2, 6

[49] Haoran Song, Wenchao Ding, Yuxuan Chen, Shaojie Shen, Michael Yu Wang, and Qifeng Chen. Pip: Planning-informed trajectory prediction for autonomous driving. In *ECCV*, pages 598–614. Springer, 2020. 1

[50] Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. In *ICCV*, pages 9601–9611, 2023. 1

[51] Luca Anthony Thiede and Pratik Prabhanjan Brahma. Analyzing the variety loss in the context of probabilistic trajectory prediction. In *ICCV*, pages 9954–9963, 2019. 1

[52] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*, pages 835–851. Springer, 2016. 1

[53] Chengxin Wang, Shaofeng Cai, and Gary Tan. Graphtcn: Spatio-temporal interaction modeling for human trajectory prediction. In *WACV*, pages 3450–3459, 2021. 2, 3

[54] Erica Weng, Hana Hoshino, Deva Ramanan, and Kris Kitani. Joint metrics matter: A better standard for trajectory forecasting. In *ICCV*, pages 20315–20326, 2023. 2, 6, 7

[55] Conghao Wong, Beihao Xia, Ziming Hong, Qinmu Peng, Wei Yuan, Qiong Cao, Yibo Yang, and Xinge You. View vertically: A hierarchical network for trajectory prediction via fourier spectrums. In *ECCV*, pages 682–700, 2020. 1, 6, 7

[56] Songpengcheng Xia, Lei Chu, Ling Pei, Jiarui Yang, Wenxian Yu, and Robert C Qiu. Timestamp-supervised wearable-based activity segmentation and recognition with contrastive learning and order-preserving optimal transport. *TMC*, 23 (12):10734–10751, 2024. 1

[57] Songpengcheng Xia, Yu Zhang, Zhuo Su, Xiaozheng Zheng, Zheng Lv, Guidong Wang, Yongjie Zhang, Qi Wu, Lei Chu, and Ling Pei. Envposer: Environment-aware realistic human motion estimation from sparse observations with uncertainty modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1839–1849, 2025. 1

[58] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *CVPR*, pages 6498–6507, 2022. 6

[59] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *CVPR*, pages 6488–6497, 2022. 6, 7

[60] 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 *CVPR*, pages 1410–1420, 2023. 6

[61] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *ECCV*, pages 511–528. Springer, 2022. 1, 6

[62] Hao Xue, Flora Salim, Yongli Ren, and Nuria Oliver. Mobtcast: Leveraging auxiliary trajectory forecasting for human mobility prediction. *NeurIPS*, 34:30380–30391, 2021. 3

[63] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020. 2

[64] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *ICCV*, pages 9813–9823, 2021. 3, 6

[65] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, pages 16010–16021, 2023. 2

[66] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1889–1899, 2024. 1

[67] Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. Multi-agent tensor fusion for contextual trajectory prediction. In *CVPR*, pages 12126–12134, 2019. 1

[68] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *CVPR*, pages 17863–17873, 2023. 2