

TOTP: Transferable Online Pedestrian Trajectory Prediction with Temporal-Adaptive Mamba Latent Diffusion

Ziyang Ren, Ping Wei*, Shangqi Deng, Haowen Tang, Jiapeng Li, Huan Li
National Key Laboratory of Human-Machine Hybrid Augmented Intelligence
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

rzyrzy@stu.xjtu.edu.cn, pingwei@xjtu.edu.cn

Abstract

Pedestrian trajectory prediction is crucial for many intelligent tasks. While existing methods predict future trajectories from fixed-frame historical observations, they are limited by the observational perspective and the need for extensive historical information, resulting in prediction delays and inflexible generalization in real-time systems. In this paper, we propose a novel task called Transferable Online Pedestrian Trajectory Prediction (TOTP), which synchronously predicts future trajectories with variable observations and enables effective task transfer under different observation constraints. To advance TOTP modeling, we propose a Temporal-Adaptive Mamba Latent Diffusion (TAMLD) model. It utilizes the Social-Implicit Mamba Synthesizer to extract motion states with social interaction and refine temporal representations through Temporal-Aware Distillation. A Trend-Conditional Mamba Decomposer generates the motion latent distribution of the future motion trends and predicts future motion trajectories through sampling decomposition. We utilize Motion-Latent Mamba Diffusion to reconstruct the latent space disturbed by imbalanced temporal noise. Our method achieves state-of-the-art results on multiple datasets and tasks, showcasing temporal adaptability and generalization ability.

1. Introduction

Pedestrian trajectory prediction aims to predict the future spatial positions of movement trajectories with observed trajectories [37], which is widely applied in autonomous driving [4, 51], social robots [24], and video surveillance [46]. Existing pedestrian trajectory prediction tasks can be broadly categorized into two types: traditional 8-12 protocol [1, 12, 28, 53, 55] and momentary prediction [29, 43].

The traditional 8-12 protocol involves extracting 8-frame observed trajectories to predict 12-frame future trajectories,

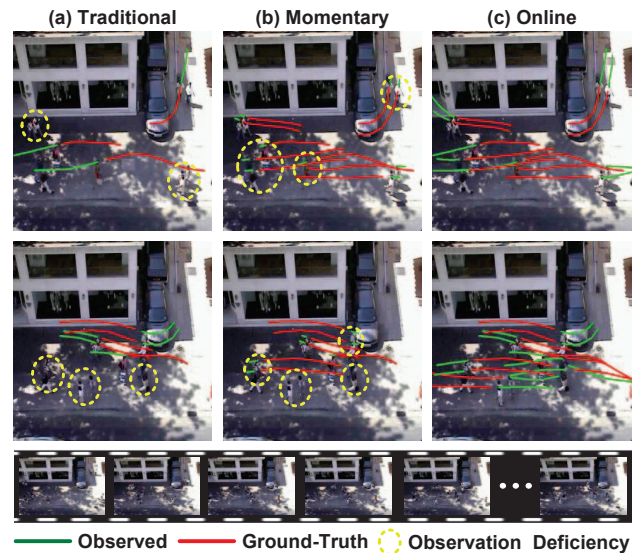


Figure 1. Comparison of three trajectory prediction tasks. Existing tasks (a) and (b) are constrained by the fixed-frame protocol, with observation omissions and insufficiencies in the limited perspective. Our proposed task (c) performs the comprehensive prediction for all observations in the video.

as shown in Fig. 1 (a). Studies [3, 26, 27, 33, 54] have achieved significant prediction accuracy in the fixed observation duration paradigm. However, due to the spatial constraints of the observation perspective, pedestrians visible for only a limited number of frames are not suitable for traditional prediction. This leads to safety risks that are selectively overlooked in practical applications. The momentary prediction reduces prediction costs and uses only 2-frame observed trajectories for future trajectory prediction, as shown in Fig. 1 (b). The simplification of this process leads to performance degradation due to the neglect of long-term information during prediction.

Although these methods have achieved impressive performance, existing tasks struggle to be applied to real-time systems due to the following three challenges.

- **Observation Constraints.** Due to view constraints, ne-

*Corresponding author.

glecting inadequate observations or observed long-term information leads to prediction delays or inadequacies.

- **Evaluation Omissions.** The exclusion of future trajectories with fewer than 12 frames from evaluation results in low utilization of the annotated data.
- **Inflexible Generalization.** The fixed observation of motion information results in ineffective transfer across tasks with different temporal information.

To address these issues, we propose the Transferable Online Pedestrian Trajectory Prediction task (TOTP). The objective of TOTP is to **perform predictions with accessible observations of different lengths for all pedestrians exhibiting movement trends within videos**, as shown in Fig. 1 (c). Meanwhile, it requires efficient transferability across traditional, momentary, and online tasks, thereby achieving seamless integration and simultaneous application in both online and offline prediction systems. With this design, the task can effectively address the diverse pedestrian prediction problems in complex scenarios, including pedestrians entering the scene, pedestrians located at the observation center, and pedestrians who are about to leave the scene simultaneously. This makes the task more challenging and of greater practical significance.

To address the requirements of TOTP, we propose a Temporal-Adaptive Mamba Latent Diffusion (TAMLD) framework, which integrates multiscale observation durations with a Mamba-based motion state space model [10]. This framework is designed to extract motion states from multi-scale trajectories, mitigate temporal sparsity imbalance, and capture real-time interactions among pedestrians with heterogeneous historical information. Specifically, we introduce a Social-Implicit Mamba Synthesizer to extract the real-time motion state reflecting implicit interactions among multiple pedestrians. Additionally, we develop a Temporal-Aware Distillation module to ensure more robust and adaptive motion dynamics by effectively reducing temporal imbalance. The motion state features are then fed into the proposed Trend-Conditional Mamba Decomposer to generate motion latent distributions that reflect future motion trends. Finally, a Motion-Latent Mamba Diffusion module is introduced to reconstruct the stable latent representations, thereby effectively alleviating the impact of imbalanced temporal noise.

Our approach primarily has three contributions:

- As far as we know, we are the first to propose Transferable Online Pedestrian Trajectory Prediction (TOTP). It overcomes the constraints of imbalanced temporal information and fixed observations in the three challenges.
- We propose Temporal-Adaptive Mamba Latent Diffusion (TAMLD) for the TOTP task. TAMLD leverages a motion state space model to extract motion states. It further generates motion latent distributions to capture motion trends and employs a diffusion model to reconstruct

future motion states.

- TAMLD achieves state-of-the-art results on two public datasets and three pedestrian trajectory prediction tasks. It effectively enables the proposed TOTP task with temporal adaptability and task transferability.

2. Related Work

2.1. Traditional Pedestrian Trajectory Prediction

Early research employs mathematical models such as Gaussian process regression [17, 50] and the social force model [13]. Subsequent studies transform trajectory prediction into a sequential processing task, utilizing Recurrent Neural Networks [1, 45, 57, 61] to extract deep features. Recent studies introduced Graph Neural Networks [16, 28, 52] and Transformer [9, 58, 59] to capture interactions among multiple individuals, gradually forming the encoder-decoder structure of the Sequence-to-Sequence paradigm. These approaches have increased reliance on trajectory information, making them susceptible to data-driven influences and difficult to transfer to other trajectory prediction tasks.

Considering the stochastic nature of pedestrian movement, deterministic trajectory prediction struggles to capture the diversity of motion modes. Therefore, some studies have introduced Generative Adversarial Network (GANs) [2, 7, 12, 15, 38] and Conditional Variational Autoencoder (CVAE) [25, 30, 39, 49, 55, 60] to model future motion as a stochastic space, and respectively approximate the known conditional motion distribution through adversarial training and the Evidence Lower Bound (ELBO). However, a single distribution still finds it challenging to represent the ambiguous complexity of motion. MID [11] sets the stochastic space as the result of overlaying random noise from multiple complex factors and combines diffusion [14] with the Transformer model [47] for multi-step reverse denoising to reconstruct the distribution of motion trends. Subsequent studies [3, 27] follow this direction, focusing on the initialization process of the stochastic space. Different from existing methods, we employ a motion state space model to extract the motion state and reconstruct the deep motion states by denoising them in a high-dimensional space.

2.2. Momentary Pedestrian Trajectory Prediction

The traditional trajectory prediction task is not suitable for real-time systems. Some studies [23, 44] propose video-based trajectory prediction to reduce data annotation costs and enhance the generalization performance. On the other hand, studies [29, 43] address sudden safety concerns by simplifying the task to predict trajectories instantly using only two-frame observed trajectories. BCDiff [21] employs a bidirectional diffusion model to reconstruct observed information and future states of instantaneous trajectories. The prediction of instantaneous pedestrian trajectories fo-

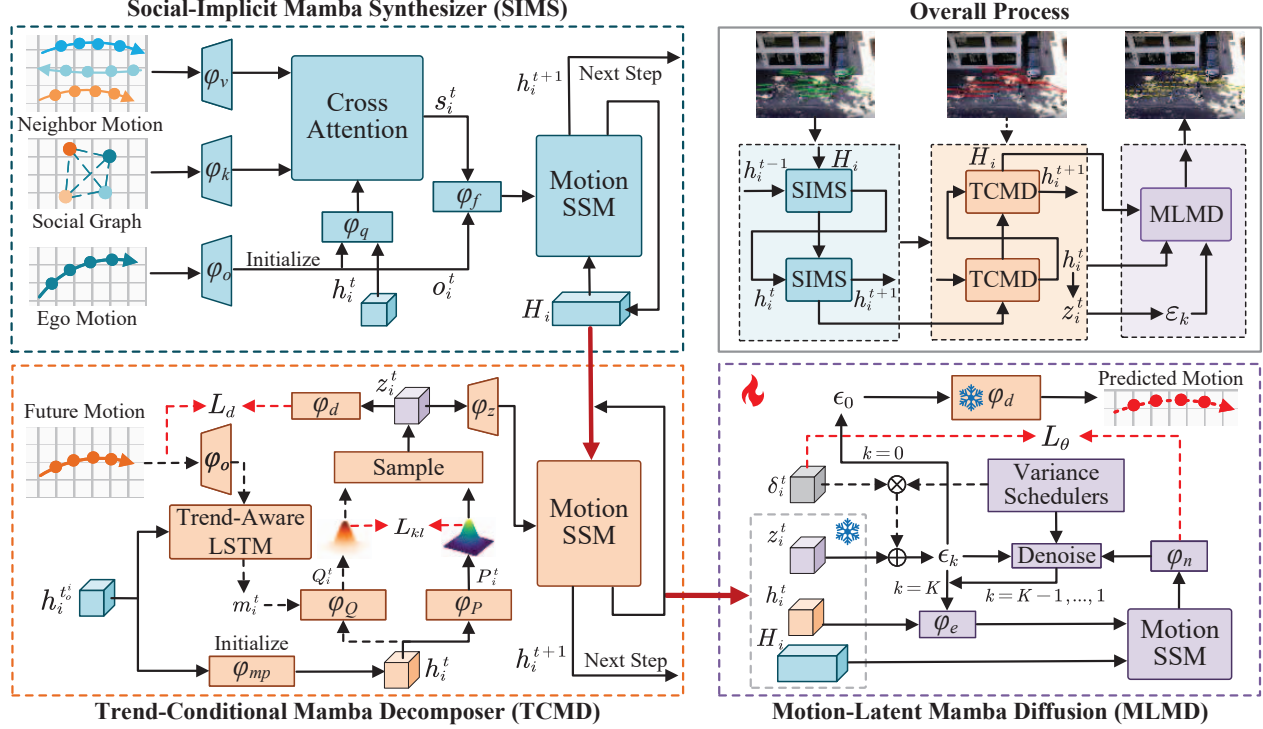


Figure 2. Architecture of Temporal-Adaptive Mamba Latent Diffusion. Dashed arrows are implemented during training.

cuses on sparse temporal motion states, yet motion changes also depend on accumulated long-term information. Our approach simultaneously analyzes motion state changes under variable-length observations in real-time videos, possessing scalability and efficient transferability tailored for imbalanced temporal settings.

2.3. Online Pedestrian Trajectory Prediction

Existing studies [32, 48] implement real-time object tracking and validate the effectiveness of the online system. Research [8] predicts vehicle online trajectories, and the approach focuses on the centralization of the agent rather than limited observations. LaKD [22] and FLN [56] propose length-agnostic vehicle trajectory prediction to overcome the temporal imbalance. SingularTrajectory [3] effectively applies the singular value space to multiple trajectory prediction tasks, yet it remains constrained by the fixed-observation protocol. Taking into account the strengths of these studies and the challenges of existing tasks, we introduce the TOTP task. This task aligns with the distribution of observations in the videos. Furthermore, our approach emphasizes the efficient transferability across multiple tasks, which includes the scalability towards sparse or ample observations driven by imbalanced temporal information.

3. Methodology

The primary objective of TOTP is to predict the future motion trajectories of each pedestrian appearing in the video

as they exhibit motion tendencies with variable observations. Specifically, we propose Temporal-Adaptive Mamba Latent Diffusion (TAML), and the overall architecture is shown in Fig. 2. It includes the Social-Implicit Mamba Synthesizer (SIMS), which extracts motion state features from the observed trajectory $x_i = \{x_i^1, x_i^2, \dots, x_i^{t_o}\}$ of the i -th pedestrian with t_o^i frames ($t_o^i \geq 2$, manifestation of the motion trend). Notably, our approach incorporates Temporal-Aware Distillation to rectify feature disparities arising from temporal sparsity. Within the Trend-Conditional Mamba Decomposer (TCMD), we generate the motion latent features constrained by future motion trends and decompose it to predict the future trajectory $y_i = \{x_i^{t_o+1}, x_i^{t_o+2}, \dots, x_i^{t_o+t_f}\}$ of the next t_f^i frames (if $t_f^i > 12$, we predict only 12 frames). And we employ Motion-Latent Mamba Diffusion (MLMD) for denoising the imbalanced temporal noise in the motion latent features.

3.1. Social-Implicit Mamba Synthesizer

We propose the motion state space model (Motion SSM) to retain the temporal information of different-scale trajectories in the motion latent state and regulate preservation with Mamba [10]. The motion latent state is influenced by interactions with other pedestrians. Therefore, we divide the input of the Motion SSM into ego-motion features and social-interaction features $\{[o_i^t, s_i^t] | t = 1, 2, \dots, t_o^i\}$. o_i^t is obtained through embedding spatial position and motion

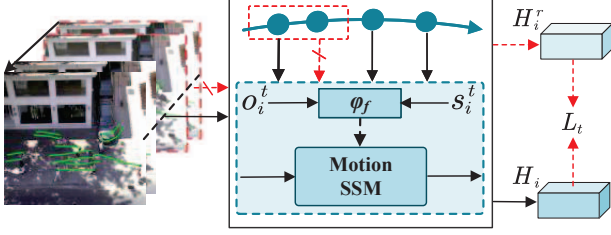


Figure 3. Structure of Temporal-Aware Distillation.

velocity, as shown in Eq. (1).

$$o_i^t = \varphi_o([x_i^t, v_i^t]), \quad v_i^t = x_i^t - x_i^{t-1}. \quad (1)$$

φ_o denotes the function of the ego-motion embedding. Following research [2], we describe the social feature k_i^t as the Euclidean distance, cosine value, and predicted distance for the next step between pedestrian i and neighbor motion $[x_{\mathbb{N}}^t, v_{\mathbb{N}}^t] = \{[x_j^t, v_j^t] | j \in \text{neighbor}(i)\}$ at time t . s_i^t is obtained through cross attention computation [47] between motion state features h_i^t and k_i^t , as shown in Eq. (2).

$$s_i^t = \text{softmax}(\varphi_q(h_i^t) \cdot \varphi_k(k_i^t)) \varphi_v([x_{\mathbb{N}}^t, v_{\mathbb{N}}^t]), \quad (2)$$

where h_i^0 is initially set to o_i^0 . φ_q , φ_k and φ_v represent the embedding functions in the cross attention mechanism.

We inject the fusion features into the motion state space model to incrementally update the transition of motion latent states, thereby synthesizing social-implicit motion state features. This process can be represented as Eq. (3).

$$h_i^{t+1}, H_i = \text{SSM}(\varphi_f([o_i^t, s_i^t]), H_i), \quad (3)$$

where φ_f and SSM denote the function of the fusion embedding and the motion state space model. The motion latent state feature H_i is initialized using the embedding of o_i^0 . The operation of SSM follows research [10].

Temporal-Aware Distillation. When extracting the real-time motion state from online trajectories, the selective mechanism can control the retention of motion latent states, stabilizing them within a certain range as they are iteratively updated. Additionally, considering the distinctiveness in observation trajectory lengths among different pedestrians, we propose Temporal-Aware Distillation to reduce the temporal information imbalance. Specifically, we randomly discard the initial t_r^i frames of the i -th pedestrian with long observed trajectories to approximate the motion latent states under temporal sparsity observations, as indicated by the red arrows in Fig. 3. The distillation loss L_t can be described by Eq. (4). \mathbb{E} denotes the average expectation.

$$L_t = \mathbb{E}_i \|H_i(x_i) - H_i^r(x_i^{t_o^i - t_r^i}, \dots, x_i^{t_o^i})\|_2^2, \quad (4)$$

where H_i^r represents the motion latent state features extracted with temporal dropout. Reducing the sparsity difference can enhance the temporal adaptability of social-implicit motion state features for synchronously predicting trajectories with observations of different scales.

3.2. Trend-Conditional Mamba Decomposer

In order to transition from real-time motion states of observed trajectories to future motion states, we propose a CVAE-based [41] motion latent generation model. We encode the ego-motion features of future motion trajectories through Trend-Aware LSTM into motion trend features m_i^t , which is shown in Eq. (5). It is utilized as a conditional constraint for the motion latent distribution.

$$\{m_i^t\} = \text{TAL}(\{o_i^t\}, h_i^{t_o^i}), \quad t = t_o^i + 1, \dots, t_o^i + t_f^i, \quad (5)$$

where TAL denotes the function of the Trend-Aware LSTM. The choice of using LSTM for the embedding of motion trends instead of Mamba is to reduce complexity, preventing overfitting that may arise from the difficulty of fitting the motion latent distribution under conditional constraints. Subsequently, $h_i^{t_o^i}$ is transformed into initial motion state features $h_i^{t_o^i+1}$ through motion state initial embedding φ_{mp} . We utilize h_i^t to generate two Gaussian distributions. The trend-conditional latent distribution Q_i^t is employed for training, while the motion latent distribution P_i^t is used for inference. The process is shown as Eq. (6).

$$Q_i^t = \varphi_Q([h_i^t, m_i^t]), \quad P_i^t = \varphi_P(h_i^t), \\ L_{kl} = \mathbb{E}_{i,t} \log\left(\frac{Q_i^t(z_i^t | h_i^t, m_i^t)}{P_i^t(z_i^t | h_i^t)}\right), \quad t = t_o^i + 1, \dots, t_o^i + t_f^i, \quad (6)$$

where φ_Q and φ_P represent the motion latent embedding functions. The motion latent feature z_i^t is sampled from the distribution Q_i^t . We approximate trend-conditional generation by minimizing the KL divergence loss L_{kl} computed based on probability $Q_i^t(z_i^t | h_i^t, m_i^t)$ and $P_i^t(z_i^t | h_i^t)$. The predicted motion velocity \hat{v}_i^t is derived by decomposing the fused features of z_i^t and h_i^t , and used to predict future motion trajectories \hat{y}_i through the accumulation function \sum_d . The decomposition process is illustrated in Eq. (7).

$$\hat{v}_i^t = \varphi_d(h_i^t, z_i^t), \quad L_d = \mathbb{E}_i \|y_i - \hat{y}_i\|_2^2, \quad \hat{y}_i = \sum_d \hat{v}_i^t, \quad (7)$$

where φ_d and L_d denotes the decoding function and the decomposed loss, respectively. We re-encode \hat{v}_i^t through embedding φ_z and inject it into the Motion SSM to selectively control the preservation of the latent state, enabling the extraction of states to adapt to different temporal scales. The process is illustrated in Eq. (8).

$$h_i^{t+1}, H_i = \text{SSM}(\varphi_z(\hat{v}_i^t, z_i^t), H_i), \quad (8)$$

where SSM follows the same computation as Eq. (3), but without parameter sharing. The motion state features and coarse-grained predicted trajectories are obtained with the iterative loop of motion state synthesis and predicted motion velocity decomposition.

3.3. Motion-Latent Mamba Diffusion

Although the motion latent distribution under the supervision of the KL divergence loss approximates the trend-conditional distribution, there still exists some discrepancy due to the influence of sampling randomness. We model the motion latent distribution as a result of the motion trend influenced by imbalanced temporal noise, and introduce the diffusion model [14] to reconstruct the motion latent space through forward noise addition and reverse denoising. Inspired by research [36], we denoise the latent space instead of the motion velocity. This aims at reducing the impact of motion variation with changes in the dataset on one hand and enhancing the dimension of motion latent representations on the other. Specifically, while freezing the parameters of previous modules, Gaussian noise δ_i^t is progressively added to z_i^t . The noised motion latent space ϵ_k is constructed as done in research [11, 34], and the backward stepwise denoising process can be represented as Eq. (9).

$$\epsilon_{k-1} = \frac{1}{\sqrt{1-\beta_k}} \left(\epsilon_k - \frac{\beta_k}{\sqrt{1-\alpha_k}} \hat{\delta}_i^t \right) + \sqrt{\beta_k} \delta_z, \quad \delta_z \in \mathcal{N}(0, 1), \quad (9)$$

where β_k denotes variance schedulers and $\alpha_k = \prod_{s=1}^k (1 - \beta_s)$. We employ the Motion SSM to obtain the predicted Gaussian noise $\hat{\delta}_i^t$ under conditions of motion state features. It accommodates the learned motion states h_i^t, H_i . This process is illustrated in Eq. (10).

$$\hat{\delta}_i^t = \varphi_n(\text{SSM}(\varphi_e(h_i^t, \epsilon_k), H_i)), \quad (10)$$

where φ_n and φ_e represent the imbalanced temporal noise decoding function and the motion latent conditional fusion function, respectively. Meanwhile, SSM does not share parameters with the previous models. During inference, the denoised motion latent feature is used to replace z_i^t in Eq. (7), and decomposed to obtain fine-grained velocities.

3.4. Training Objective

The training process of our proposed model is divided into two phases.

Phase 1: Motion State Extraction and Motion Latent Generation. During this phase, our objective is to extract effective motion state features, enhance the generation of motion latent features with trend conditions, and reduce the differences in temporal sparsity and approximate probability distribution. The loss function formulation can be represented as Eq. (11), where $\gamma_1, \gamma_2, \gamma_3$ are the hyperparameters controlling the loss weights.

$$L = \gamma_1 L_d + \gamma_2 L_{kl} + \gamma_3 L_t. \quad (11)$$

Phase 2: Denoising Reconstruction of Motion Latent Space. While freezing the parameters learned in the previous phase, our objective is to minimize the prediction loss of imbalanced temporal noise. Therefore, $L_\theta = \mathbb{E}_{i,t} \|\delta_i^t - \hat{\delta}_i^t\|_2^2$ is used for supervised training in this phase.

4. Experiments

4.1. Experimental Setup

Dataset. ETH [31]-UCY [20] and Stanford Drone Dataset (SDD) [35] are two widely used real-world pedestrian trajectory datasets. ETH-UCY collectively contains over 1500 pedestrians from five scenes (ETH, HOTEL, UNIV, ZARA1, ZARA2). We follow the leave-one-out evaluation protocol, where four sub-datasets are used for training, and the remaining one is used for testing. SDD includes pedestrian motion trajectory data from a large number of campus scenarios from bird’s-eye view perspectives. Due to the limited field of view in the original video, pedestrians who have just entered, those with sufficient movement history, and those about to leave can coexist simultaneously. Therefore, online trajectory prediction is crucial for real-time observation systems in these datasets.

Evaluation Metrics. Referring to the widely used Average Displacement Error (ADE) and Final Displacement Error (FDE) metrics, we introduce evaluation metrics ADE_o and FDE_o for variable-length online pedestrian trajectory prediction. Concurrently, we follow the Best-of-N protocol in research [12, 55], which involves generating 20 prediction trajectories simultaneously and evaluating the best result.

$$ADE_o = \mathbb{E}_{i,t} \|y_i^t - \hat{y}_i^t\|_2, \quad FDE_o = \mathbb{E}_i \|y_i^{t_f} - \hat{y}_i^{t_f}\|_2. \quad (12)$$

Implementation Details. The structure of TAML D consists of the Mamba [10], CVAE [41], and Diffusion model [14]. The feature dimension is set to 256, the State feature dimension for the Motion State Space Model is set to 512, and the motion latent dimension is set to 32. In the first phase, the Adam optimizer [18] is utilized with hyperparameters $\gamma_1, \gamma_2, \gamma_3$ set to 1, 0.5, 0.5 for training 500 epochs. The second phase involves training for 200 epochs.

4.2. Quantitative Analysis

We analyze the performance of TOTP in two aspects: online pedestrian trajectory prediction and task transfer prediction.

Online Pedestrian Trajectory Prediction. We compare our method with existing state-of-the-art (SOTA) methods on the online trajectory prediction task, as shown in Table 1. The results indicate that TAML D achieves state-of-the-art online trajectory prediction results on all datasets. On the ETH-UCY dataset, our approach has significantly improved by 44.7% on ADE_o and 43.5% on FDE_o compared to the latest SOTA method [19]. In comparison to the RNN-based method [55] and Transformer-based method [40], TAML D also shows improvements of 27.6%/22.2% and 51.2%/53.3% on ADE_o/FDE_o , which is attributed to Mamba’s compatibility with controlling motion states of varying-duration trajectories. On the SDD dataset, our approach demonstrates a 5.87%/4.51% performance advantage in ADE_o/FDE_o metrics. Through comparison, it elu-

Dataset	PECNet [25]	CausalHTP [5]	LB-EBM [30]	MemoNet [53]	SocialVAE [55]	EqMotion [54]	TUTR [40]	MART [19]	TAMLD
ETH	0.94/1.47	0.82/1.32	0.75/1.17	0.58/1.04	<u>0.44/0.70</u>	0.69/1.14	0.63/1.07	0.59/0.93	0.35/0.56
HOTEL	0.55/0.87	0.56/0.97	0.44/0.78	0.29/0.55	<u>0.22/0.28</u>	0.32/0.67	0.26/0.45	0.25/0.41	0.13/0.21
UNIV	0.72/0.96	0.59/1.01	0.55/0.96	0.32/0.71	<u>0.30/0.52</u>	0.45/0.91	0.47/0.86	0.42/0.75	0.24/0.44
ZARA1	0.86/1.14	0.66/1.09	0.64/0.96	0.31/0.59	<u>0.24/0.37</u>	0.40/0.74	0.42/0.71	0.32/0.52	0.16/0.29
ZARA2	0.77/0.94	0.59/0.93	0.55/0.91	0.30/0.57	<u>0.23/0.36</u>	0.38/0.71	0.37/0.66	0.30/0.48	0.15/0.26
AVG	0.77/1.08	0.64/1.06	0.59/0.96	0.36/0.69	<u>0.29/0.45</u>	0.45/0.83	0.43/0.75	0.38/0.62	0.21/0.35
SDD	10.16/16.22	9.14/15.93	8.97/15.90	8.55/12.44	8.37/ <u>12.26</u>	8.79/12.95	8.58/12.82	<u>8.35/12.43</u>	7.86/11.87

Table 1. The ADE_o/FDE_o metric comparison of the online trajectory prediction task on the ETH-UCY (meters) and SDD (pixels) datasets. The lower the metric, the better the prediction performance. The **bold/underlined** font represents the best/second-best results.

Dataset	Social-GAN [12]	PECNet [25]	LB-EBM [30]	MemoNet [53]	GroupNet [52]	EqMotion [54]	TUTR [40]	MART [19]	TAMLD
ETH	0.87/1.62	0.54/0.87	0.30/0.52	0.40/0.61	0.46/0.73	0.40/0.61	0.40/0.61	<u>0.35/0.47</u>	0.39/0.58
HOTEL	0.67/1.37	0.18/0.24	0.13/0.20	0.11/0.17	0.15/0.25	0.12/ <u>0.18</u>	0.11/0.18	0.14/0.22	0.13/ <u>0.18</u>
UNIV	0.76/1.52	0.35/0.60	0.27/0.52	0.24/0.43	0.26/0.49	<u>0.23/0.43</u>	<u>0.23/0.42</u>	0.25/0.45	0.22/0.37
ZARA1	0.35/0.68	0.22/0.39	0.20/0.37	<u>0.18/0.32</u>	0.21/0.39	<u>0.18/0.32</u>	<u>0.18/0.34</u>	0.17/0.29	<u>0.18/0.28</u>
ZARA2	0.42/0.84	0.17/0.30	0.15/0.29	0.14/0.24	0.17/0.33	0.13/0.23	0.13/0.25	0.13/0.22	0.13/0.21
AVG	0.61/1.21	0.29/0.48	0.21/0.38	0.21/0.35	0.25/0.44	0.21/0.35	0.21/0.36	0.21/0.33	0.21/0.32

Table 2. The ADE/FDE metric (meters) comparison of the traditional trajectory prediction task on the ETH-UCY datasets.

cidates that existing methods rely on the similarity of the trajectory data, while our approach can extract motion state features from varying-scale online trajectories efficiently.

Traditional Pedestrian Trajectory Prediction. It is worth noting that TAMLD’s prediction performance on the online trajectory prediction task closely approaches that of the existing traditional prediction task. To further validate the adaptability of our approach in pedestrian trajectory analysis, we evaluate the prediction performance of the ETH-UCY dataset on the traditional trajectory prediction task, as shown in Table 2. Our method demonstrates high scalability and achieves SOTA results in the traditional prediction task. Compared to MART [19], our approach demonstrates improvements of 7.14%/18.18% and 12.0%/17.8% on HOTEL and UNIV datasets, which include diverse motion modalities and complex social interactions.

Momentary Pedestrian Trajectory Prediction. We also conduct momentary trajectory prediction, and the comparison result is shown in Table 3. Compared to existing research that experiences significant performance drops due to sparse information, our method maintains relatively stable prediction results. Additionally, TAMLD achieves an improvement of 4.00%/5.00% compared to the SOTA method [3]. Even when dealing with sparse temporal information, TAMLD still effectively extracts momentary motion states for accurately predicting future motion trajectories, which further validates its temporal adaptability.

Task Transfer Prediction. Unlike existing research that focuses on fixed observation tasks, our approach aims to build a unified framework that not only adapts to temporal information but also synchronously performs multiple trajectory

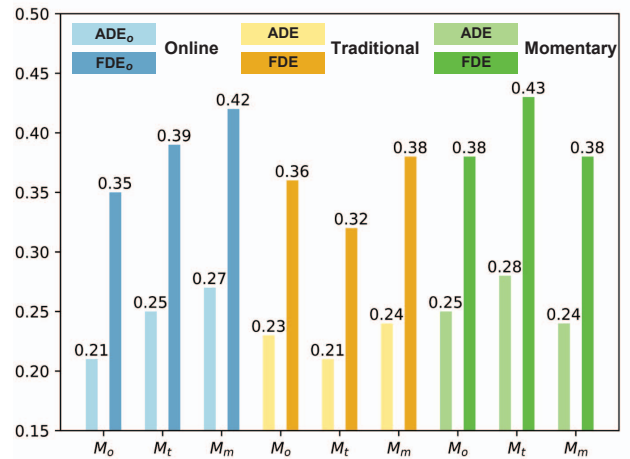


Figure 4. Task transfer prediction performance.

prediction tasks. **Leveraging the temporal adaptability of TAMLD, models trained on any task can be directly transferred to the other two tasks.** The average prediction results of models trained on the three tasks and transferred to the remaining tasks are shown in Fig. 4, where M_o , M_t , M_m represent models trained under online, traditional, and momentary task settings. M_t and M_m trained on traditional and momentary tasks that rely on the similarity of data can effectively transfer to other tasks, albeit with some performance reduction. Furthermore, M_o demonstrates comparable performance transferred to other tasks. It especially showcases performance in momentary trajectory prediction that closely approximates its performance under the corresponding momentary prediction. This indirectly elaborates that online pedestrian trajectory prediction

Dataset	PECNet [25]	AgentFormer[59]	MID [11]	SocialVAE [55] +BCDiff [21]	STT+DTO [29]	MOE-Traj++ [43]	SingularTrajectory [3]	TAMLD
ETH	0.64/1.04	1.10/2.11	0.63/1.05	0.53/0.91	0.62/1.22	0.64/1.12	0.45/0.67	0.45/0.66
HOTEL	0.28/0.53	0.50/1.02	0.29/0.49	<u>0.17/0.27</u>	0.29/0.56	0.20/0.33	0.18/0.29	0.15/0.23
UNIV	0.28/0.49	0.52/1.10	0.30/0.56	<u>0.24/0.40</u>	0.58/1.14	0.33/0.62	<u>0.24/0.43</u>	0.23/0.41
ZARA1	0.25/0.44	0.56/1.18	0.30/0.56	0.21/0.37	0.45/0.98	0.22/0.42	0.19/0.33	<u>0.20/0.33</u>
ZARA2	0.19/0.34	0.43/0.89	0.22/0.40	0.16/0.26	0.34/0.74	0.17/0.32	0.17/0.28	0.16/0.26
AVG	0.33/0.57	0.62/1.26	0.35/0.61	0.26/0.44	0.46/0.93	0.31/0.56	<u>0.25/0.40</u>	0.24/0.38

Table 3. The ADE/FDE metric (meters) comparison of the momentary trajectory prediction task on the ETH-UCY datasets.

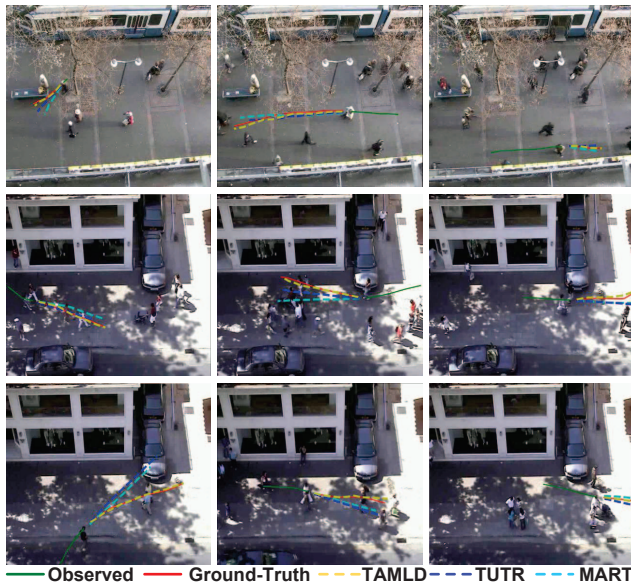


Figure 5. Visualization of the predicted trajectory.

can enhance the scalability of our method. These results indicate the generalization ability of our approach.

4.3. Qualitative Analysis

To highlight the accuracy of our prediction results, the visualization of predicted trajectories is shown in Fig. 5 and compared with the results of TUTR [40] and MART [19]. From left to right, we sequentially present three instances of observed initial, central, and final states. Whether the observed temporal information is sparse or ample in historical data, our approach consistently yields optimal predicted results. The motion patterns predicted by the other two methods exhibit monotonicity under the influence of temporal imbalance. TAMLD can effectively extract the diversity of motion patterns and accurately predict future motion trajectories in various scenarios, including straight-line motion, turns, and sudden changes in direction.

While exhibiting temporal adaptability, our approach also maintains task scalability. The visualization of predicted trajectories for three tasks using models trained under three different task settings is shown in Fig. 6. In any task setting, TAMLD can be effectively applied to other tasks with precise prediction results. For trajectories with

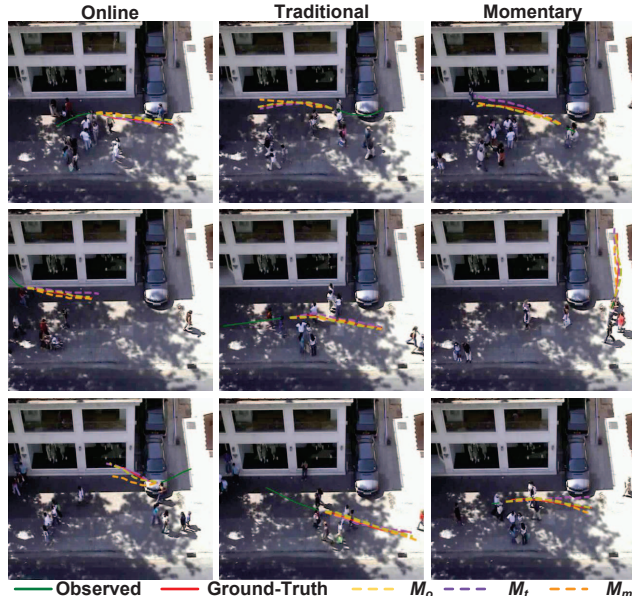


Figure 6. Visualization of Task Transfer Prediction.

limited observational data, M_m can extract momentary motion states. M_t retains long-term information and accurately predicts complex motion patterns of the pedestrians at the center of observation. Importantly, M_o exhibits comparable prediction performance in all tasks. In conclusion, the online trajectory prediction task facilitates effective transfer across tasks. Our method demonstrates both task scalability and temporal adaptability.

4.4. Ablation Study

To elucidate the effectiveness of each module proposed by us in extracting motion states and generating latent motion, we conduct extensive ablation studies of the online task on the HOTEL, ZARA1, and ZARA2 datasets. These datasets incorporate multiple motion patterns, social interactions, and complex temporal information. The results of the ablation study are shown in Table 4. The removal modules include social interaction features s , Temporal-Aware Distillation (TAD), Trend-Aware LSTM (TAL), and Motion-Latent Mamba Diffusion (MLMD). All proposed modules are effective, with s being utilized to capture multiple pedestrian interactions. Although TAD does not no-

s	TAD	TAL	MLMD	HOTEL	ZARA1	ZARA2
-	-	-	-	0.17/0.24	0.26/0.37	0.23/0.34
✓	-	-	-	0.18/0.25	0.24/0.35	0.22/0.32
-	✓	-	-	0.17/0.24	0.24/0.36	0.22/0.33
✓	✓	-	-	0.17/0.24	0.22/0.34	0.21/0.32
✓	✓	✓	-	0.15/0.23	0.19/0.31	0.18/0.29
✓	✓	✓	✓	0.13/0.21	0.16/0.29	0.15/0.26

Table 4. Ablation study results.

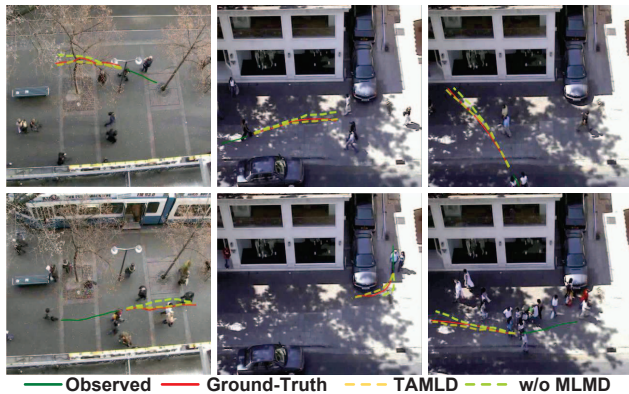


Figure 7. Visualization of the ablation study for MLMD.

ticeably improve the predictions, it contributes to the task transfer efficiency and the temporal adaptability of extracted motion states. TAL utilizes motion trends to enhance the continuity of future motion latent generation. MLMD reduces the instability in latent sampling caused by unknown conditions, thus both modules contribute to improving prediction performance.

We also conduct visual comparisons of MLMD, as illustrated in Fig. 7. MLMD reduces the imbalanced temporal noise of the motion latent features and ensures stability. It can yield more stable results and accurately predict future motion trajectories even with fewer observations and complex motion variations. The number of denoising steps and the denoising method impact the reconstruction process of the motion latent space. Therefore, we conduct an ablation study of the denoising process, as shown in Table 5. Due to employing KL divergence loss to approximate the motion latent distribution constrained by the motion trend condition, we require fewer denoising steps to achieve effective reconstruction results. Additionally, DDPM refines denoising details to improve prediction results.

Our model is built upon the Mamba model [10] and exhibits performance improvements compared to existing RNN-based and Transformer-based methods. To further demonstrate the effectiveness of the proposed Motion SSM, we conduct ablation studies by replacing it with GRU [6] and Transformer [47]. The comparison results, parameters (M), and FLOPs (G) are shown in Fig. 8. The Best-of-N protocol requires generating 20 prediction trajectories simultaneously for evaluation, and Motion SSM still demonstrates significant performance improvements. GRU

Method	Steps	HOTEL	ZARA1	ZARA2
DDPM	6	0.14/0.22	0.17/0.30	0.15/0.26
	10	0.13/0.21	0.16/0.29	0.15/0.26
	20	0.14/0.23	0.17/0.30	0.17/0.27
	40	0.15/0.24	0.18/0.30	0.18/0.28
DDIM	6	0.15/0.24	0.18/0.29	0.16/0.27
	10	0.14/0.22	0.17/0.29	0.16/0.27
	20	0.15/0.23	0.17/0.30	0.17/0.27
	40	0.15/0.24	0.18/0.31	0.17/0.28

Table 5. Ablation study of DDPM [14] and DDIM [42].

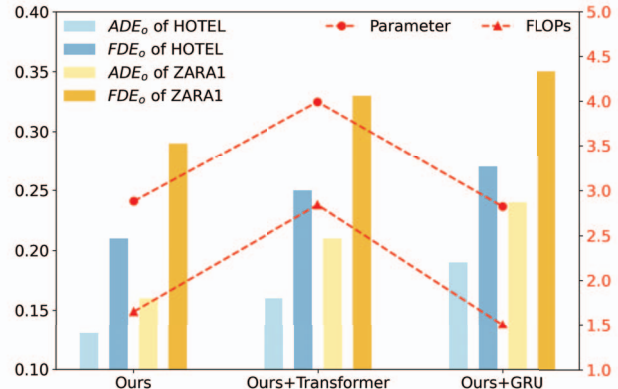


Figure 8. Ablation study comparison of the Mamba model.

has larger errors in prediction results. The Transformer improves the prediction performance but has a large number of parameters and a slower inference speed. Motion SSM can accurately predict future motion trajectories with less computational cost. These results demonstrate that the Mamba model can be effectively applied to online trajectory prediction with imbalanced temporal observations.

5. Conclusion

This paper proposes a novel task, called TOTP, to predict the real-time trajectory of all pedestrians in videos under different observation constraints and efficient transfer across multiple tasks. Building upon this, we construct TAMLD, which utilizes the Mamba model to extract real-time motion states and generate the motion latent feature under future trend conditional constraints. Considering the influence of imbalanced temporal noise, we integrate the diffusion model to reconstruct the motion latent space. Our method achieves SOTA results on multiple tasks and exhibits scalability in both the temporal and task domains.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No.62088102, No. U23B2060) and National Key Laboratory Under Grant 241-HF-D10-01.

References

- [1] Alexandre Alahi, Kratharth 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
- [2] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *CVPRW*, pages 0–0, 2019. 2, 4
- [3] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singular trajectory: Universal trajectory predictor using diffusion model. In *CVPR*, pages 17890–17901, 2024. 1, 2, 3, 6, 7
- [4] Haoyu Bai, Shaojun Cai, Nan Ye, David Hsu, and Wee Sun Lee. Intention-aware online pomdp planning for autonomous driving in a crowd. In *IEEE international Conference on Robotics and Automation*, pages 454–460, 2015. 1
- [5] Guangyi Chen, Junlong Li, Jiwen Lu, and Jie Zhou. Human trajectory prediction via counterfactual analysis. In *ICCV*, pages 9824–9833, 2021. 6
- [6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 8
- [7] Patrick Dendorfer, Sven Elflein, and Laura Leal-Taixé. M-gan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction. In *ICCV*, pages 13158–13167, 2021. 2
- [8] Wenchao Ding and Shaojie Shen. Online vehicle trajectory prediction using policy anticipation network and optimization-based context reasoning. In *International Conference on Robotics and Automation*, pages 9610–9616, 2019. 3
- [9] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. In *ICPR*, pages 10335–10342, 2021. 2
- [10] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2, 3, 4, 5, 8
- [11] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *CVPR*, pages 17113–17122, 2022. 2, 5, 7
- [12] 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. 1, 2, 5, 6
- [13] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 2
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 5, 8
- [15] Yue Hu, Siheng Chen, Ya Zhang, and Xiao Gu. Collaborative motion prediction via neural motion message passing. In *CVPR*, pages 6319–6328, 2020. 2
- [16] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *ICCV*, pages 6272–6281, 2019. 2
- [17] Kihwan Kim, Dongryeol Lee, and Irfan Essa. Gaussian process regression flow for analysis of motion trajectories. In *ICCV*, pages 1164–1171, 2011. 2
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [19] Seongju Lee, Junseok Lee, Yeonguk Yu, Taeri Kim, and Kyoobin Lee. Mart: Multiscale relational transformer networks for multi-agent trajectory prediction. *arXiv preprint arXiv:2407.21635*, 2024. 5, 6, 7
- [20] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, pages 655–664, 2007. 5
- [21] Rongqing Li, Changsheng Li, Dongchun Ren, Guangyi Chen, Ye Yuan, and Guoren Wang. Bcdiff: Bidirectional consistent diffusion for instantaneous trajectory prediction. *Advances in Neural Information Processing Systems*, 36: 14400–14413, 2023. 2, 7
- [22] Yuhang Li, Changsheng Li, Ruilin Lv, Rongqing Li, Ye Yuan, and Guoren Wang. Lakd: Length-agnostic knowledge distillation for trajectory prediction with any length observations. In *Neural Information Processing Systems*, 2024. 3
- [23] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *CVPR*, pages 11553–11562, 2020. 2
- [24] Matthias Luber, Johannes A Stork, Gian Diego Tipaldi, and Kai O Arras. People tracking with human motion predictions from social forces. In *IEEE international Conference on Robotics and Automation*, pages 464–469, 2010. 1
- [25] 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. 2, 6, 7
- [26] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *ICCV*, pages 15233–15242, 2021. 1
- [27] 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, 2
- [28] Abdullah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *CVPR*, pages 14424–14432, 2020. 1, 2
- [29] Alessio Monti, Angelo Porrello, Simone Calderara, Pasquale Coscia, Lamberto Ballan, and Rita Cucchiara. How many observations are enough? knowledge distillation for trajectory forecasting. In *CVPR*, pages 6553–6562, 2022. 1, 2, 7
- [30] 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, 6

- [31] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *CVPR*, pages 261–268, 2009. 5
- [32] J Redmon. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 3
- [33] Ziyang Ren, Ping Wei, Haowen Tang, Huan Li, and Jin Yang. Learning scene-goal-aware motion representation for trajectory prediction. In *British Machine Vision Conference*, 2024. 1
- [34] Ziyang Ren, Ping Wei, Haowen Tang, Huan Li, Jin Yang, and Jialu Qin. Stochastic-aware mamba diffusion for pedestrian trajectory prediction. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025. 5
- [35] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, pages 549–565, 2016. 5
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 5
- [37] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrilu, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020. 1
- [38] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezafofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *CVPR*, pages 1349–1358, 2019. 2
- [39] Tim Salzman, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *ECCV*, pages 683–700, 2020. 2
- [40] Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *ICCV*, pages 9675–9684, 2023. 5, 6, 7
- [41] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, 28:3483–3491, 2015. 4, 5
- [42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 8
- [43] Jianhua Sun, Yuxuan Li, Liang Chai, Hao-Shu Fang, Yong-Lu Li, and Cewu Lu. Human trajectory prediction with momentary observation. In *CVPR*, pages 6467–6476, 2022. 1, 2, 7
- [44] Haowen Tang, Ping Wei, Huan Li, Jiapeng Li, and Nanning Zheng. Relation reasoning for video pedestrian trajectory prediction. In *ICME*, pages 1–6, 2022. 2
- [45] Chaofan Tao, Qinhong Jiang, Lixin Duan, and Ping Luo. Dynamic and static context-aware lstm for multi-agent motion prediction. In *ECCV*, pages 547–563, 2020. 2
- [46] Maria Valera and Sergio A Velastin. Intelligent distributed surveillance systems: a review. *IEE Proceedings-Vision, Image and Signal Processing*, 152(2):192–204, 2005. 1
- [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 4, 8
- [48] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024. 3
- [49] Chuhua Wang, Yuchen Wang, Mingze Xu, and David J. Crandall. Stepwise goal-driven networks for trajectory prediction. *IEEE Robotics and Automation Letters*, 7(2):2716–2723, 2022. 2
- [50] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE TPAMI*, 30(2):283–298, 2008. 2
- [51] Weishang Wu, Xiaoheng Deng, Ping Jiang, Shaohua Wan, and Yuanxiong Guo. Crossfuser: Multi-modal feature fusion for end-to-end autonomous driving under unseen weather conditions. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 1
- [52] 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. 2, 6
- [53] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: retrospective-memory-based trajectory prediction. In *CVPR*, pages 6488–6497, 2022. 1, 6
- [54] 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. 1, 6
- [55] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *ECCV*, pages 511–528, 2022. 1, 2, 5, 6, 7
- [56] Yi Xu and Yun Fu. Adapting to length shift: Flexilength network for trajectory prediction. In *CVPR*, pages 15226–15237, 2024. 3
- [57] Hao Xue, Du Q. Huynh, and Mark Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1186–1194, 2018. 2
- [58] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *ECCV*, pages 507–523, 2020. 2
- [59] Ye Yuan, Xinchao Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *ICCV*, pages 9813–9823, 2021. 2, 7
- [60] Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory prediction via neural social physics. In *ECCV*, pages 376–394, 2022. 2
- [61] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *CVPR*, pages 12085–12094, 2019. 2