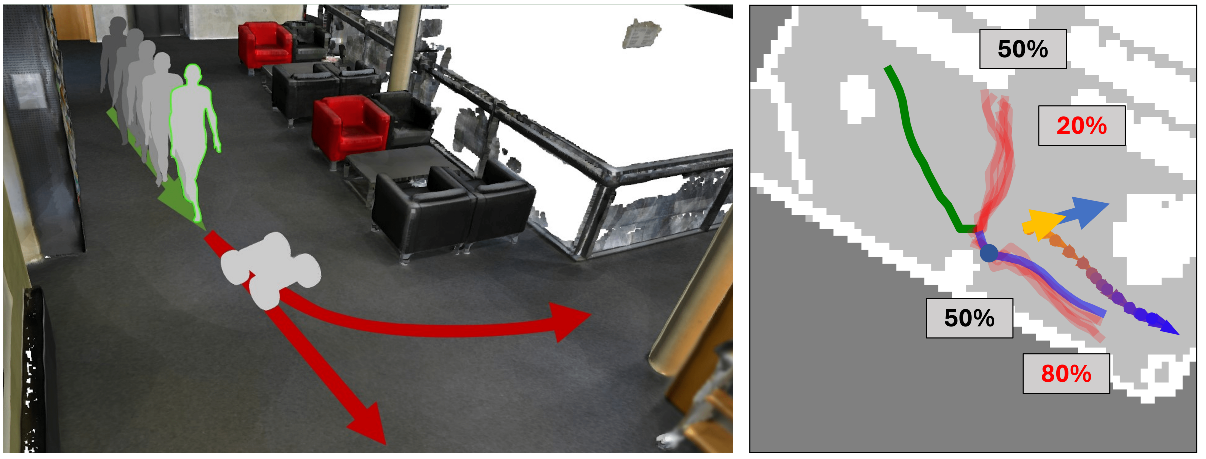
Closed-loop Diffusion Planning over Multi-Modal Distributions for Robot Follow-Ahead

Author Names Omitted for Anonymous Review.



(a) **Robot Follow-Ahead at a junction** (b) **Illustration: Closed-loop diffusion planning**

Fig. 1: **Motivation scenario for closed-loop diffusion planning (CLDP)**. (a) A challenging instance of the robot follow-ahead task where a human is entering a junction and the robot needs to stay in front of the human. (b) We propose to solve such scenarios by 1) predicting a set of human motion samples using a diffusion model, visualized in red lines, and 2) plan our robot path (dots with arrows in fading colors), with a given initial pose (the yellow arrow), by adjusting the weight for each sample with posterior probability. The background is an occupancy map, visualized as white (1) for obstacles, grey (0) for free space, and dark (-1) for unknown space. The past and future (ground truth) human motion are drawn in green and blue lines on the map.

Abstract— **Diffusion models are increasingly used in robotics applications due to their ability to represent high-dimensional, multi-modal distributions. However, sampling and inference using diffusion models are time-consuming. Therefore, it is hard to use them for real-time planning in dynamic scenarios. In this paper, we focus on such a scenario: the robot follow-ahead (RFA) task, where the robot's objective is to maintain its relative position in front of a moving human actor while keeping the actor in view. It is a challenging task because when a human approaches a junction, the robot needs to predict the human's trajectory in advance and plan accordingly, without knowing the exact branch of the junction the human will take. To address the dynamic nature of this task while utilizing the expressive power of diffusion models, we present a recursive Bayesian Filter whose initial prior distribution is generated by a learned diffusion process conditioned on the observed trajectory. To enable fast planning and to incorporate sensor measurements, we perform Bayesian updates using a local motion model, until the next estimate from the diffusion process arrives. Once we have an accurate estimate of the human trajectory distribution, we show that the optimal robot motion strategy for the given horizon can be computed based on the estimated distribution. Experiments are conducted on multiple datasets to evaluate our model's ability to represent the trajectory distribution and the performance of the closed-loop diffusion planning strategy in achieving the robot follow-ahead task. They indicate that our closed-loop diffusion planning strategy outperforms baseline planning strategies and is more responsive to human motion. We also demonstrate the algorithm in an indoor RFA task on a real robot. Project page and supplementary materials can be found at https://cldp-rfa.com**

I. Introduction

Diffusion models form a class of generative models that have been successfully applied to image generation [1]–[3], and more recently to motion planning [4]–[8]. They have the advantage of modeling high-dimensional multi-modal distributions [4] and are made "controllable" using reward functions that guide the sampling process, for example, to avoid collisions or reach waypoints [8]–[10]. However, such sampling methods are time-consuming. While there exist methods that try to speed up the inference process [2], [11], [12], they either sacrifice the quality of the samples or remain computationally expensive. This sampling efficiency issue makes it difficult to apply diffusion models to scenarios with fast temporal dynamics. Robot follow-ahead (RFA) [13]–[15] is one such scenario with applications such as auto-cinematography [16], [17] or cargo carrying [18]. A mobile robot is asked to "follow" the target human while staying in front of them to maintain visibility. However, human motion can be highly dynamic. When a human approaches a junction in an indoor environment, without knowing which direction the human will enter, the robot needs to be reactive when the human selects the exact branch of the junction (Fig. 1.(a)). This is challenging for classical open-loop diffusion-based planning strategies.

Existing works [13], [14], [19] try to solve this RFA challenge by modeling the uncertainty of the future human motion with a single, mostly-likely trajectory. Their underlying assumption is that the future motion distribution can be represented as a high-dimensional Gaussian, and these methods report the mean of the Gaussian as the prediction. In the junction scenario, however, the underlying distribution is multi-modal, and representing it as a Gaussian leads to predictions that may not even be feasible. Therefore, to precisely describe the future human motion distribution, we propose to model multi-modal distributions using diffusion models. Instead of generating a single motion prediction at a time, we aim to predict the distribution of all possible future trajectories.

Moreover, to achieve real-time performance in dynamic scenarios, we present a recursive Bayesian Filter that estimates the future human trajectory distribution in two stages. 1) It generates the initial prior distribution from a diffusion model conditioned on the past human trajectory and the surrounding environment. 2) Then, it observes human motion and efficiently updates the distribution in a closed form using the Bayesian rule. We adjust our plan for the robot path based on distribution updates without having to wait for another prediction.

We illustrate this two-stage method with a motivating example in Fig. 1.(b) : An actor is walking along the green line in an indoor environment and is about to enter a junction. Suppose the actor has equal chances of going to the up and right corridors. For now, let's assume that we know this prior precisely. The best planning strategy for the robot (beginning at the yellow arrow) is to "remain centered" (blue arrow) until the next prediction result is received. In this paper, instead of waiting for the next prediction, we propose to close the planning loop by observing the actor's motion (blue line). Based on the observed actor's location (blue dot), we can adjust the predicted distribution (e.g., 20% versus 80%) and reconfigure the robot's path before the next motion prediction is generated. By doing so, we can leverage the long-term information from the prediction while remaining responsive to the observed human motion. In this paper, we present a method (**CLDP**) that can reason about and react to such complex human motions in a dynamic manner. Specifically:

1) We show how to generate a faithful distribution of future human trajectories using samples from a diffusion model.
2) We present a recursive Bayesian Filter method on top of the diffusion model to update the distribution estimation in a closed form and plan the robot path efficiently.
3) We conduct experiments to evaluate our method on both simulation and real-world datasets. We show that our diffusion model can precisely represent the future human trajectory distribution, and our closed-form diffusion planning algorithm can plan the robot path responsively and achieve the robot follow-ahead task well.

II. Related Work

In this section, we summarize related literature along two main aspects: 1) diffusion models used in robot motion planning, 2) human motion forecasting (synthesis) for robot follow-ahead.

a) Diffusion for Robot Planning: Generative models based on diffusion processes have been successful for image generation [2], [3], [20]–[22]. They have been recently adapted to robot motion-planning tasks [7], [8], in robot manipulation [4], [6], [23] and navigation [12], [23], [24] tasks. These methods use diffusion for their ability to represent multi-modal distributions [4], [24], or their controllability [9], [23] by adding cost terms to avoid collision, to ensure planning goals, or to solve trajectory optimization problems [25], [26]. However, all these methods follow a plan-and-execute scheme and focus on low-dynamic scenarios. The inference time is a bottleneck in highly-dynamic situations such as our robot follow-ahead problem. Therefore, we propose to decouple the objective prediction process using the diffusion models and planning processes so that our planner can be more agile and reactive to the objective.

b) Human Motion Prediction for Robot Follow-Ahead: The robot follow-ahead has been studied since [27], [28] by estimating the human future motion with Kalman-Filter-based probabilistic models [27], [29], [30] or single prediction deep learning neural networks [13], [14]. Besides, there are series of study that focus on the human motion prediction (synthesis), potentially incorporate environmental information [31]–[36], using MLP [37], GANs [38]–[40], Graph Convolutional Networks (GCN) [41], cVAE [31], [33], and Transformers [42]–[45], or most recently diffusion models [23], [46], [47]. In this paper, we build on top of the state-of-the-art diffusion methods [4], [23] for predicting future human motion and focus on designing a closed-loop Bayesian filter for the planning algorithm.

III. Closed-loop Diffusion Planning

In this section, we formulate the robot-following-ahead problem by dividing it into three sub-problems. 1) Human motion prediction (Sec. III-A). 2) Robot path planning given the predicted human trajectories (Sec. III-B). 3) Closed-loop planning (Sec. III-C).

A. Human Motion Prediction

Given the past human motion and the surrounding environment, we formulate the human motion prediction problem as follows. Denote the human 2D trajectory history as $\tau_{1:t_0}$ (simplified as $\tau_{:t_0}$) , and the future trajectory as $\tau_{t_0:t_0+T}$ (simplified as $\tau_{t_0:}$) with time horizon time horizon T. Given the human motion history $\tau_{:t_0}$ and surrounding environment S, we calculate the conditional future human trajectory probability $p(\tau_{t_0:}|\tau_{:t_0}, S)$. Then, we sample N trajectories from the probability, denoted as $\{\hat{\tau}^i_{t_0:} \,|\, i = [1, ..., N]\}$), to represent the future human motion distribution. $\hat{\tau}^i_{t_0:} \sim p(\tau_{t_0:}|\tau_{:t_0}, S)$

We parameterize this conditioned probability $p_\phi(\tau_{t_0:}|\tau_{:t_0}, S)$ with a trained diffusion model ϕ. Specifically, we use DDPM [1] as our sampling method. We condition our sampling on the past trajectory $\tau_{:t_0}$ with the 2D occupancy map S by encoding them into a latent feature z_{cond}. Note that the map is centered at the current human's pose, with its heading direction aligned with the x-axis of the map. We implement two state-of-the-art neural network architectures to learn the conditioned noise $\epsilon_\phi(\tau^{(j)}, t, z_{cond})$. **Dif-TR** from [23] is based on spatial transformer [45]. **Dif-Unet** from [4] uses 1-D convolution U-net [48] structure. Since the network design is not our main contribution, we include the network architecture figures and implementation details in Supplementary II.A. In this formulation, it is assumed that the training dataset faithfully represents the human's trajectory for the particular application.

B. Path Planning for Robot Follow-ahead

Given the human motion prediction set $\{\hat{\tau}_{t_0:}\}$, we formulate our path planning problem as a finite-horizon optimization problem.

a) Robot Model: We use the unicycle model for the robot. We define the robot state $\mathbf{X}_t$ at time t by its 2D position x, y and its yaw angle θ. $\mathbf{X} = (x, y, \theta)$. We define the robot control command as velocity and rotation $\mathbf{u} = (v, \omega)$. Given a sequence of control inputs $\mathcal{U} = \{\mathbf{u}_{t_0}, \mathbf{u}_{t_1}, \dots, \mathbf{u}_{t_{T-1}}\}$, we have the future robot states $\mathcal{X}_{t_0:}(\mathcal{U}) = \{\mathbf{X}_{t_1}, \mathbf{X}_{t_2}, \dots, \mathbf{X}_{t_T}\}$ calculated as Eq. 1.

$$\mathbf{X}_{t+1} = \mathbf{X}_t + \mathbf{B}\mathbf{u}\Delta t, \quad \mathbf{B} = \begin{bmatrix} \cos(\theta) & 0 \\ \sin(\theta) & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} v \\ \omega \end{bmatrix} \tag{1}$$

b) Robot Motion Planning: The cost function for a given human trajectory $\hat{\tau}^i_{t_0:}$ and a robot trajectory $\mathcal{X}_{t_0:}$ is denoted by $\mathcal{J}\left(\hat{\tau}^i_{t_0:}, \mathcal{X}_{t_0:}(\mathcal{U})\right)$. Our formulation can be adapted to arbitrary cost functions as long as it is additive over time. For our RFA problem, the cost function is defined using the viewing quality metric $\mathcal{L}_v$: the negative value of Pixels-Per-Area values from [49]; In addition, we add a cost component to constrain the robot's path so as to avoid collisions with the environment. The details of the overall cost function, given below, are presented in Supplementary I.

$$\mathcal{J}\left(\hat{\tau}^i_{t_0:}, \mathcal{X}_{t_0:}(\mathcal{U})\right) = \sum_t \gamma^t \left(\lambda_v \mathcal{L}_v + \lambda_{col}\mathcal{L}_{col}\right) \tag{2}$$

The planning task can then be formulated as a finite-horizon optimization problem of minimizing the expected cost over the given human motion distribution, i.e., given the cost function $\mathcal{J}$, we calculate a sequence of robot control inputs $\mathcal{U}$, such that the expected cost function is minimized, which can be approximated using the sample mean (Eq. 3).

$$\begin{aligned} \mathcal{U} &= \arg\min_{\mathcal{U}} \mathop{\mathbb{E}}_{\tau \sim p(\tau_{t_0:}|\tau_{:t_0}, S)} \left[\mathcal{J}(\tau, \mathcal{X}_{t_0:}(\mathcal{U}))\right] \\ &= \arg\min_{\mathcal{U}} \int p(\tau|\tau_{:t_0}, S)\mathcal{J}(\tau, \mathcal{X}_{t_0:}(\mathcal{U}))d\tau \\ &\approx \arg\min_{\mathcal{U}} \frac{1}{N} \sum_{\tau \in \{\hat{\tau}_{t_0:}\}} \mathcal{J}(\tau, \mathcal{X}_{t_0:}(\mathcal{U})) \end{aligned} \tag{3}$$

C. Closed-loop Planning

Given a set of predicted human motion trajectory samples, solving Eq. 3 will give us an open-loop solution $\mathcal{U}$. Our motivating scenario (Fig. 1) shows that this open-loop solution can be suboptimal – as it may choose to 'stay in the middle' at junctions. To close the planning loop, we treat the predicted probability as a prior and update it with the posterior probability based on human motion observations.

a) Posterior Probability: Suppose now we observe k-steps human motion $\tau_{t_0:t_k}$. The posterior probability can be calculated with the Bayesian Equation as in Eq. 4,

$$p(\hat{\tau}_{t_0:}|\tau_{t_0:t_k}, \tau_{:t_0}, S) \propto p(\hat{\tau}_{t_0:}|\tau_{:t_0}, S) \cdot p(\tau_{t_0:t_k}|\hat{\tau}_{t_0:}, \tau_{:t_0}, S) \tag{4}$$

The posterior probability mainly consists of two terms. 1) the prior probability for each predicted sample, and 2) the probability of our observation conditioned on the state, which is related to our human localization method. To define the probability across the entire trajectory space, we assume a) motion estimation is irrelevant to the surrounding environment and the history of motion, $p(\tau_{t_0:t_k}|\hat{\tau}^i_{t_0:}, \tau_{:t_0}, S) = p(\tau_{t_0:t_k}|\hat{\tau}^i_{t_0:})$; b) *locally* the probability of the observation conditioned on the state is independent over time and the likelihood is a Gaussian for each time step. As in Eq. 5, we define a probability distribution conditioned on a base trajectory as follows:

$$\begin{aligned} p(\tau_{t_0:t_k}|\hat{\tau}^i_{t_0:}) &= \prod_{t=t_0,\dots,t_k} p(\tau_t|\hat{\tau}^i_t) \\ &\propto \exp\left(\sum_{t=t_0,\dots,t_k} -\frac{1}{2\sigma_t^2}\|\tau_t - \hat{\tau}^i_t\|^2\right) \end{aligned} \tag{5}$$

Note that σ_t is a hyperparameter defining the covariance (temperature parameter). Since the human position distribution would have a higher variance over time, we can define $\sigma_t = \eta^t\sigma_0$ in practice, where η is a growth factor for the covariance change.

b) Closed-loop planning: Given k-step human motion observations, we re-plan the robot path by minimizing the expectation of the cost function over a **posterior probability** as in Eq. 6. However, since we only have samples $\tau \in \{\hat{\tau}_{t_0:}\}$ that are sampled following the prior distribution, we need to update the weight for each sample, using the *Importance Sampling* technique [50], and approximate the expectation with a weighted cost over all samples.

$$\begin{aligned}
&\underset{\tau\sim p(\hat{\tau}_{t_0:}|\tau_{t_0:t_k},\tau_{:t_0},S)}{\mathbb{E}}\left[\mathcal{J}(\tau,\mathcal{X}_{t_0:}(\mathcal{U}))\right]\\
&=\int p(\tau|\tau_{t_0:t_k},\tau_{:t_0},S)\mathcal{J}(\tau,\mathcal{X}_{t_0:}(\mathcal{U}))d\tau\\
&=\int p(\tau|\tau_{:t_0},S)\frac{p(\tau|\tau_{t_0:t_k},\tau_{:t_0},S)}{p(\tau|\tau_{:t_0},S)}\mathcal{J}(\tau,\mathcal{X}_{t_0:}(\mathcal{U}))d\tau\\
&\approx\frac{1}{N}\sum_{\tau\in\{\hat{\tau}_{t_0:}\}}\frac{p(\tau|\tau_{t_0:t_k},\tau_{:t_0},S)}{p(\tau|\tau_{:t_0},S)}\mathcal{J}(\tau,\mathcal{X}_{t_0:}(\mathcal{U}))
\end{aligned}\tag{6}$$

Once we apply the Bayesian equation in Eq. 4 and the sensor model in Eq. 5, we can plan the robot motion in a closed-loop fashion by updating the posterior probability and solving Eq. 7.

$$\begin{aligned}
\mathcal{U}&=\arg\min_{\mathcal{U}}\underset{\tau\sim p(\hat{\tau}_{t_0:}|\tau_{t_0:t_k},\tau_{:t_0},S)}{\mathbb{E}}\left[\mathcal{J}(\tau,\mathcal{X}_{t_0:}(\mathcal{U}))\right]\\
&\approx\arg\min_{\mathcal{U}}\frac{1}{N}\sum_{\tau\in\{\hat{\tau}_{t_0:}\}}\frac{p(\tau|\tau_{t_0:t_k},\tau_{:t_0},S)}{p(\tau|\tau_{:t_0},S)}\mathcal{J}(\tau,\mathcal{X}_{t_0:}(\mathcal{U}))\\
&=\arg\min_{\mathcal{U}}\frac{1}{N}\sum_{\tau\in\{\hat{\tau}_{t_0:}\}}p(\tau_{t_0:t_k}|\hat{\tau}^i_{t_0:})\mathcal{J}(\tau,\mathcal{X}_{t_0:}(\mathcal{U}))\\
&=\arg\min_{\mathcal{U}}\frac{1}{N}\sum_{\tau\in\{\hat{\tau}_{t_0:}\}}\lambda_i\mathcal{J}(\tau,\mathcal{X}_{t_0:}(\mathcal{U}))
\end{aligned}\tag{7}$$

To solve this optimization problem, we use a dynamic programming method to calculate an optimal solution or MPPI [51]/log-MPPI [52] for a sub-optimal solution. We repeatedly update the importance weight for each sample λ_i and plan the robot path until we get the next prediction sample set. We include further planning details with an algorithm pseudo-code in Supplementary II.C. We also discuss the advantages of such an algorithm in robotic system design in Supplementary II.C.

IV. Experiments

In this section, we conduct experiments to answer the following two questions. 1) How well does the diffusion model represent the future human trajectory distribution? 2) How does our closed-loop planning strategy benefit from the posterior observation and improve the planning results? We will first introduce our experiment setup (dataset) in Sec. IV-A, then provide our analysis to both questions in Sec. IV-B and Sec. IV-C. We provide a real robot demonstration in Sec. IV-C.0.d, also in Supplementary III.D.

A. Datasets

This paper mainly uses two datasets, **GTA-IM** [31] (simulation) and **HPS** [53] (real-world), to evaluate our method. Both contain complete 3D environment point clouds and human 3D skeleton poses in a building-scale area. Additionally, our investigation of these datasets reveals that both datasets lack the diversity of human motion in some scenarios. Therefore, we create a dummy dataset **Junc** (short for "Junction") where the actor turns at the perpendicular T-junctions at a random position (sampled uniformly from an interval in its front) and at a binary direction (either turn left or right) with equal (50%) probability. We include more details on the dataset in the Supplementary III.A.

B. Human Motion Prediction

In our first experiment, we investigate the performance of the diffusion models in predicting human motion compared to other state-of-the-art approaches. For each method, $N=10$ future trajectories are sampled.

a) Metric: To evaluate the performance of such distribution prediction, we use Average Displacement Error **(ADE)** and Final Displacement Error **(FDE)** as our two metrics. In addition, same as [47], we use **1-minADE** and **k-minADE** to evaluate the average of 1- and k- minimum average ADE, and also **1-minFDE** and **k-minFDE** to evaluate 1- and k- minimum average FDE. In our experiment, we set $k=5$. For all these metrics, lower error means better performance.

b) Baselines: We select a few representative baselines for human motion prediction and the RFA task: **(a) PathNet** [13] and **(b) STPOTR** [14] are the state-of-the-art neural networks for predicting long-term human motion. Also, we include **(c) TR** where we use our noise prediction network in **Dif-TR** to predict the human future motion directly. These three baselines represent the classical one-prediction methods. Note that since only one prediction result is predicted, 1-minADE and k-minADE are the same as ADE, as are 1-minFDE, k-minFDE, and FDE. In the meantime, we include **(d) cVAE**, as a representative of the generative method. We modify a conditional Variational Autoencoder (cVAE) [21] with the same multi-ahead attention architecture in our **Dif-TR** method.

c) Results: We provide both qualitative (Fig. 2) and quantitative (Table. I) results. Results show that diffusion models from both implementations (**Dif-TR** and **Dif-Unet**) outperform the 1-minADE/1-minFDE on all datasets and the k-minADE/k-minFDE on GTA-IM dataset while performing worse on the ADE/FDE metric. A worse ADE and FDE are acceptable because we will close the planning loop and weigh more on the sample closest (1-minADE) to the ground-truth trajectory. Meanwhile, qualitative results highlight that the diffusion models can provide multi-model predictions in a few samples when cVAE can not. This is more significant in the dataset **Junc**, where the output is strictly bimodal. One potential explanation is that cVAE enforces the latent space to be Gaussian, whereas in some of our scenarios, this may be problematic due to the bimodal distribution. Both quantitative and qualitative results indicate that diffusion models better cover the entire trajectory space. We provide extra discussion on selecting generative models and the sampling methods in Supplementary II.B. We also provide experiments on different human representations in Supplementary III.B.

C. Closed-loop diffusion planning

Given the results of human motion prediction, we evaluate the robot's following-ahead performance among different

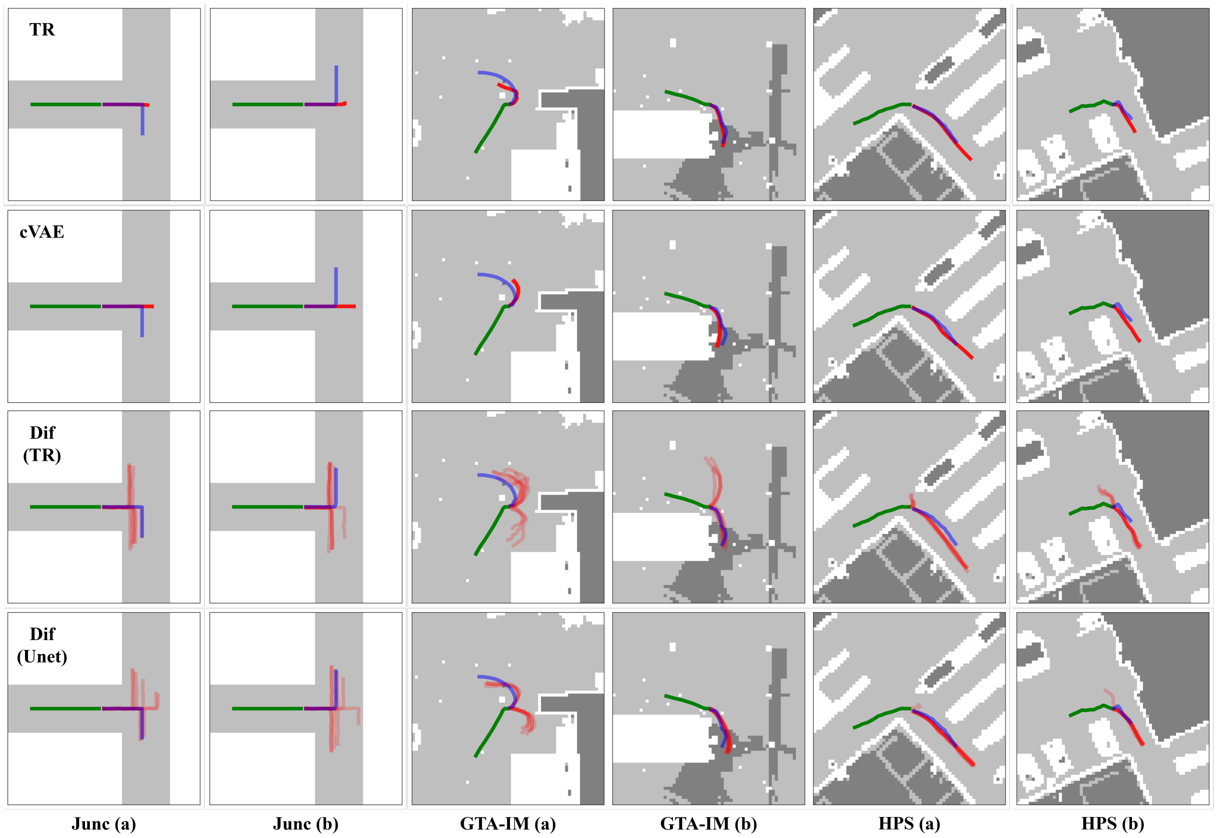


Fig. 2: **Qualitative results for human motion prediction**. We show human history trajectory (green), multiple human motion prediction samples (red), and future motion ground truth (blue) on the occupancy map. Each column is a different scenario, and each row uses a different method. We show that diffusion models perform better when modeling multi-modal distribution on human future motion.

TABLE I: **Human trajectory prediction**.

Dataset	**Junc**			**GTA-IM**			**HPS**		
Method	1-ADE	k-ADE	ADE	1-ADE	k-ADE	ADE	1-ADE	k-ADE	ADE
	1-FDE	k-FDE	FDE	1-FDE	k-FDE	FDE	1-FDE	k-FDE	FDE
PathNet [13]	0.234 0.913	0.234 0.913	0.234 0.913	0.193 0.357	0.193 0.357	0.193 0.357	0.711 1.596	0.711 1.596	0.711 1.596
STPOTR [14]	0.305 1.046	0.305 1.046	0.305 1.046	0.187 0.302	0.187 0.302	0.187 **0.302**	0.718 1.415	0.718 1.415	0.718 1.415
TR	0.518 1.393	0.518 1.393	0.518 1.393	0.179 0.349	0.179 0.349	**0.179** 0.349	0.667 1.380	**0.667** **1.380**	**0.667** **1.380**
cVAE [21]	0.193 0.403	**0.198** **0.415**	**0.204** **0.427**	0.200 0.417	0.207 0.433	0.213 0.448	0.709 1.473	0.714 1.482	0.717 1.492
Dif-TR (ours)	0.135 **0.283**	0.365 0.851	0.688 1.791	**0.136** **0.239**	**0.162** **0.303**	0.197 0.382	**0.612** **1.279**	0.693 1.450	0.773 1.613
Dif-Unet (ours)	**0.089** 0.284	0.220 0.712	0.381 1.335	0.176 0.322	0.200 0.372	0.234 0.440	0.663 1.337	0.724 1.470	0.794 1.614
Dif-TR (map only)	0.115 0.276	0.240 0.683	0.403 1.315	0.198 0.350	0.243 0.440	0.376 0.699	0.679 1.356	0.735 1.476	0.795 1.600
Dif-TR (with pose)	0.102 0.307	0.229 0.710	0.392 1.335	0.118 0.227	0.132 0.263	0.147 0.300	0.494 1.042	0.516 1.095	0.540 1.148

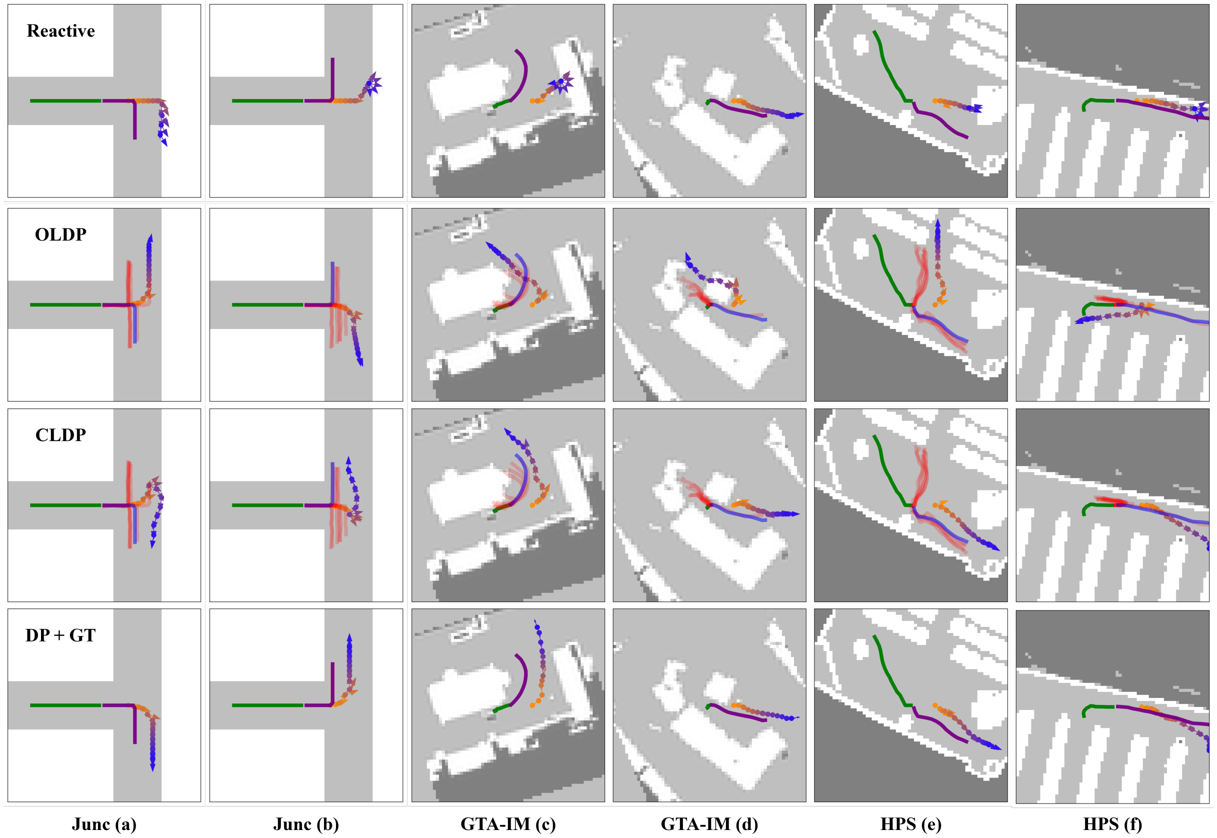


Fig. 3: **Qualitative results for different planning strategies.** We show the planned robot's 2D poses based on the predicted human trajectories in dots with arrows, in faded color from yellow to blue. Each column is a different scenario, and each row uses a different planning strategy. Our **CLDP** strategy can be responsive to human motion and keep following the human.

planning strategies. In our **CLDP** results, we mainly report the dynamic programming results for its optimality.

a) Metric: We mainly use two metrics to evaluate the planned path. We use (a) the total **cost** defined in Eq. 2 along the path and (b) the **successful rate** of the robot follow-ahead task: A *Success* is obtained if the robot maintaining in front of the human at the last time-step, and we calculate the successful rate among all data points.

b) Baselines: We compare our planning strategy against other methods, including: open-loop dynamic programming planner based on human motion prediction from (1) **PathNet** [13], (2) **STPOTR** [14], (3) open-loop diffusion planner (**OLDP**), our diffusion prediction results without posterior observation, and (4) **DP+GT**, the ground truth of future human motion trajectory. This oracle algorithm serves as the upper-bound planner for this task. These methods represent the classical planning and execution scheme. In the meantime, we also compare against: (5) a **Reactive** approach: no future motion is predicted. It demonstrates the performance of a myopic, greedy planner that directly reacts to the human position at each time step. And (6) a diffusion policy (**Dif. Policy**) approach that directly samples the robot path.

c) Results: We show planning qualitative results in Fig. 3 and quantitative results in Table II. The quantitative results in Table. II show that our **CLDP** strategy can achieve a higher success rate and a lower accumulated cost, indicating that the robot keeps track of humans better and maintains a better viewing direction. Meanwhile, we provide qualitative results in Fig. 3 as examples showing **CLDP** being able to (1) select the closest sample from a multi-modal prediction distribution during planning (Sample c, d, e), (2) recover from a wrong initial plan (Sample a, b, f) and (3) leverage the advantage of the prediction while reactive planner stuck at the local minimum (Sample b, c, e). We further provide

TABLE II: **Comparison on planning results.**

Dataset	**Junc**		**GTA-IM**		**HPS**	
Method	cost↓	succ.↑	cost↓	succ.↑	cost↓	succ.↑
DP+GT (oracle)	0.106	0.999	1.151	0.912	3.100	0.831
PathNet [13]	0.612	0.181	1.797	0.693	3.823	0.639
STPOTR [14]	0.732	0.167	1.582	0.608	3.813	0.671
OLDP	0.594	0.552	1.212	0.787	3.284	0.691
Reactive	0.561	0.268	1.598	0.703	4.991	0.593
Dif. Policy [4]	0.588	0.541	1.326	0.706	3.301	0.623
CLDP (ours)	**0.473**	**0.593**	**1.182**	**0.834**	**3.171**	**0.714**

detailed examples of the closed-loop importance weight-adjusting process in Supplementary III.C.

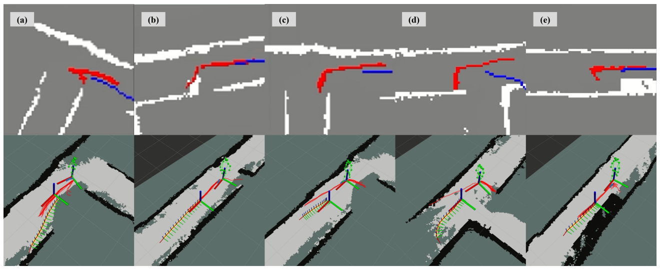


Fig. 4: **Real robot demonstration** We visualize five pairs of human motion prediction in the actor frame (top) and robot **CLDP** planning (bottom) results in 3rd-person view. Human trajectories are plotted as red lines in both figures. The planned robot path is plotted using a blue line in the top image and a yellow line with multiple frames in the bottom image. We also visualize the robot's current frame and the human skeleton pose for each sample.

d) Real Robot Demonstration: We deployed **CLDP** on a real robot system. Figure 4 shows a qualitative result for robot follow-ahead on this system. Through experiments from the real-world setting, we show that our algorithm is capable of predicting a complex distribution of the human future motion, and our ability to update the distribution and plan a robot's path on top of it. Additional details, including 1) robot hardware and system architecture; 2) human motion tracking and future motion prediction with **Dif-TR**; 3) robot **CLDP** planning with log-MPPI [52], and further qualitative results are presented in Supplementary III.D.

V. Conclusions

In this paper, we presented a recursive Bayesian Filter to inform robot planning tasks with highly-dynamic objectives, such as the robot follow-ahead (RFA) task. We divided the RFA problem into two sub-problems: 1) an initial estimation of the future human motion distribution with a diffusion model, and 2) a finite-horizon optimization problem for robot path planning with a recursive Bayesian Filter based on the posterior observation. We conducted experiments with simulation and real-world datasets to evaluate the human motion prediction and closed-loop planning results. We showed that our proposed closed-loop diffusion planning method improves planning results by utilizing the learned human motion pattern while remaining highly reactive. We also validated our **CLDP** algorithm in the real robot setting, demonstrating that we handle the RFA task well, especially in junction scenarios.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[2] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," Oct. 2022, arXiv:2010.02502 [cs]. [Online]. Available: http://arxiv.org/abs/2010.02502

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Apr. 2022, arXiv:2112.10752 [cs]. [Online]. Available: http://arxiv.org/abs/2112.10752

[4] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," Jun. 2023, arXiv:2303.04137 [cs]. [Online]. Available: http://arxiv.org/abs/2303.04137

[5] K. Mangalam, Y. An, H. Girase, and J. Malik, "From Goals, Waypoints & Paths To Long Term Human Trajectory Forecasting," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 15 213–15 222. [Online]. Available: https://ieeexplore.ieee.org/document/9709992/

[6] X. Fang, C. R. Garrett, C. Eppner, T. Lozano-Pérez, L. P. Kaelbling, and D. Fox, "DiMSam: Diffusion Models as Samplers for Task and Motion Planning under Partial Observability," Oct. 2023, arXiv:2306.13196. [Online]. Available: http://arxiv.org/abs/2306.13196

[7] S. Levine, "Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review," May 2018, arXiv:1805.00909 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1805.00909

[8] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with Diffusion for Flexible Behavior Synthesis," Dec. 2022, arXiv:2205.09991 [cs]. [Online]. Available: http://arxiv.org/abs/2205.09991

[9] J. Carvalho, A. T. Le, M. Baierl, D. Koert, and J. Peters, "Motion Planning Diffusion: Learning and Planning of Robot Motions with Diffusion Models," Aug. 2023, arXiv:2308.01557 [cs]. [Online]. Available: http://arxiv.org/abs/2308.01557

[10] M. Uehara, Y. Zhao, C. Wang, X. Li, A. Regev, S. Levine, and T. Biancalani, "Inference-Time Alignment in Diffusion Models with Reward-Guided Generation: Tutorial and Review," Jan. 2025, arXiv:2501.09685 [cs]. [Online]. Available: http://arxiv.org/abs/2501.09685

[11] K. Frans, D. Hafner, S. Levine, and P. Abbeel, "One Step Diffusion via Shortcut Models," Oct. 2024, arXiv:2410.12557. [Online]. Available: http://arxiv.org/abs/2410.12557

[12] M. Seo, Y. Cho, Y. Sung, P. Stone, Y. Zhu, and B. Kim, "PRESTO: Fast motion planning using diffusion models based on key-configuration environment representation," Sep. 2024, arXiv:2409.16012 [cs]. [Online]. Available: http://arxiv.org/abs/2409.16012

[13] Q. Jiang, B. Susam, J.-J. Chao, and V. Isler, "Map-Aware Human Pose Prediction for Robot Follow-Ahead," Mar. 2024, arXiv:2403.13294 [cs]. [Online]. Available: http://arxiv.org/abs/2403.13294

[14] M. Mahdavian, P. Nikdel, M. TaherAhmadi, and M. Chen, "STPOTR: Simultaneous Human Trajectory and Pose Prediction Using a Non-Autoregressive Transformer for Robot Following Ahead," Sep. 2022, arXiv:2209.07600 [cs]. [Online]. Available: http://arxiv.org/abs/2209.07600

[15] S. Leisiazar, E. J. Park, A. Lim, and M. Chen, "An MCTS-DRL Based Obstacle and Occlusion Avoidance Methodology in Robotic Follow-Ahead Applications," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 221–228.

[16] R. Bonatti, A. Bucker, S. Scherer, M. Mukadam, and J. Hodgins, "Batteries, camera, action! Learning a semantic control space for expressive robot cinematography," *arXiv:2011.10118 [cs]*, Mar. 2021, arXiv: 2011.10118. [Online]. Available: http://arxiv.org/abs/2011.10118

[17] T. Nägeli, L. Meier, A. Domahidi, J. Alonso-Mora, and O. Hilliges, "Real-time planning for automated multi-view drone cinematography," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 132:1–132:10, Jul. 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073712

[18] "gita robots - Piaggio Fast Forward." [Online]. Available: https://piaggiofastforward.com/

[19] T. Salzmann, H.-T. L. Chiang, M. Ryll, D. Sadigh, C. Parada, and A. Bewley, "Robots That Can See: Leveraging Human Pose for Trajectory Prediction," *IEEE Robotics and Automation Letters*, pp. 1–8, 2023, conference Name: IEEE Robotics and Automation Letters.

[20] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," Nov. 2015, arXiv:1503.03585 [cond-mat, q-bio, stat]. [Online]. Available: http://arxiv.org/abs/1503.03585

[21] K. Sohn, H. Lee, and X. Yan, "Learning Structured Output Representation using Deep Conditional Generative Models," in *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates,

Inc., 2015. [Online]. Available: https://papers.nips.cc/paper_files/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html

[22] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-Based Generative Modeling through Stochastic Differential Equations," Feb. 2021, arXiv:2011.13456 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2011.13456

[23] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu, "Diffusion-based generation, optimization, and planning in 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 750–16 761.

[24] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "NoMaD: Goal Masked Diffusion Policies for Navigation and Exploration," Oct. 2023, arXiv:2310.07896 [cs]. [Online]. Available: http://arxiv.org/abs/2310.07896

[25] C. Pan, Z. Yi, G. Shi, and G. Qu, "Model-Based Diffusion for Trajectory Optimization," May 2024, arXiv:2407.01573 [cs]. [Online]. Available: http://arxiv.org/abs/2407.01573

[26] K. Elamvazhuthi, D. Gadginmath, and F. Pasqualetti, "Denoising Diffusion-Based Control of Nonlinear Systems," Feb. 2024, arXiv:2402.02297 [math]. [Online]. Available: http://arxiv.org/abs/2402.02297

[27] D. M. Ho, J.-S. Hu, and J.-J. Wang, "Behavior control of the mobile robot for accompanying in front of a human," in *2012 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2012, pp. 377–382.

[28] N. Karnad and V. Isler, "Modeling human motion patterns for multi-robot planning," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 3161–3166.

[29] P. Nikdel, R. Shrestha, and R. Vaughan, "The Hands-Free Push-Cart: Autonomous Following in Front by Predicting User Trajectory Around Obstacles," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 4548–4554, iSSN: 2577-087X.

[30] Y. Oh, S. Choi, and S. Oh, "Chance-constrained target tracking for mobile robots," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. Seattle, WA, USA: IEEE, May 2015, pp. 409–414. [Online]. Available: http://ieeexplore.ieee.org/document/7139031/

[31] Z. Cao, H. Gao, K. Mangalam, Q.-Z. Cai, M. Vo, and J. Malik, "Long-term Human Motion Prediction with Scene Context," Jul. 2020, arXiv:2007.03672 [cs]. [Online]. Available: http://arxiv.org/abs/2007.03672

[32] W. Mao, R. I. Hartley, M. Salzmann, and others, "Contact-aware human motion forecasting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7356–7367, 2022.

[33] M. Hassan, D. Ceylan, R. Villegas, J. Saito, J. Yang, Y. Zhou, and M. Black, "Stochastic Scene-Aware Motion Prediction," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 11 354–11 364. [Online]. Available: https://ieeexplore.ieee.org/document/9710735/

[34] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black, "Resolving 3D Human Pose Ambiguities with 3D Scene Constraints," Aug. 2019, arXiv:1908.06963 [cs]. [Online]. Available: http://arxiv.org/abs/1908.06963

[35] J. Wang, Y. Rong, J. Liu, S. Yan, D. Lin, and B. Dai, "Towards Diverse and Natural Scene-aware 3D Human Motion Synthesis," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 20 428–20 437. [Online]. Available: https://ieeexplore.ieee.org/document/9880007/

[36] J. Wang, H. Xu, J. Xu, S. Liu, and X. Wang, "Synthesizing Long-Term 3D Human Motion and Interaction in 3D Scenes," Jun. 2021, arXiv:2012.05522 [cs]. [Online]. Available: http://arxiv.org/abs/2012.05522

[37] A. Bouazizi, A. Holzbock, U. Kressel, K. Dietmayer, and V. Belagiannis, "MotionMixer: MLP-based 3D Human Body Pose Forecasting," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization, Jul. 2022, pp. 791–798. [Online]. Available: https://www.ijcai.org/proceedings/2022/111

[38] B. Chopin, N. Otberdout, M. Daoudi, and A. Bartolo, "3D Skeleton-based Human Motion Prediction with Manifold-Aware GAN," Mar. 2022, arXiv:2203.00736 [cs]. [Online]. Available: http://arxiv.org/abs/2203.00736

[39] P. Nikdel, M. Mahdavian, and M. Chen, "DMMGAN: Diverse Multi Motion Prediction of 3D Human Joints using Attention-Based Generative Adversarial Network," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 9938–9944.

[40] J. Wang, S. Yan, B. Dai, and D. Lin, "Scene-aware Generative Network for Human Motion Synthesis," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 12 201–12 210. [Online]. Available: https://ieeexplore.ieee.org/document/9578765/

[41] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning Trajectory Dependencies for Human Motion Prediction," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 9488–9496. [Online]. Available: https://ieeexplore.ieee.org/document/9009559/

[42] L. Chen, R. Liu, X. Yang, D. Zhou, Q. Zhang, and X. Wei, "STTG-net: a Spatio-temporal network for human motion prediction based on transformer and graph convolution network," *Visual Computing for Industry, Biomedicine, and Art*, vol. 5, no. 1, p. 19, Dec. 2022. [Online]. Available: https://vciba.springeropen.com/articles/10.1186/s42492-022-00112-5

[43] T. Lucas, F. Baradel, P. Weinzaepfel, and G. Rogez, "PoseGPT: Quantization-based 3D Human Motion Generation and Forecasting," Oct. 2022, arXiv:2210.10542 [cs]. [Online]. Available: http://arxiv.org/abs/2210.10542

[44] A. Martinez-Gonzalez, M. Villamizar, and J.-M. Odobez, "Pose Transformers (POTR): Human Motion Prediction with Non-Autoregressive Transformers," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. Montreal, BC, Canada: IEEE, Oct. 2021, pp. 2276–2284. [Online]. Available: https://ieeexplore.ieee.org/document/9607511/

[45] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges, "A Spatio-temporal Transformer for 3D Human Motion Prediction," in *2021 International Conference on 3D Vision (3DV)*, Dec. 2021, pp. 565–574, iSSN: 2475-7888.

[46] S. Saadatnejad, A. Rasekh, M. Mofayezi, Y. Medghalchi, S. Rajabzadeh, T. Mordan, and A. Alahi, "A generic diffusion-based approach for 3D human pose prediction in the wild," Mar. 2023, arXiv:2210.05669 [cs]. [Online]. Available: http://arxiv.org/abs/2210.05669

[47] W. Wang, C. K. Liu, and M. K. III, "EgoNav: Egocentric Scene-aware Human Trajectory Prediction," Aug. 2024, arXiv:2403.19026. [Online]. Available: http://arxiv.org/abs/2403.19026

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[49] Q. Jiang and V. Isler, "Onboard View Planning of a Flying Camera for High Fidelity 3D Reconstruction of a Moving Actor," Jul. 2023, arXiv:2308.00134 [cs]. [Online]. Available: http://arxiv.org/abs/2308.00134

[50] T. Kloek and H. K. Van Dijk, "Bayesian estimates of equation system parameters: an application of integration by Monte Carlo," *Econometrica: Journal of the Econometric Society*, pp. 1–19, 1978, publisher: JSTOR.

[51] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Aggressive driving with model predictive path integral control," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1433–1440.

[52] I. S. Mohamed, K. Yin, and L. Liu, "Autonomous Navigation of AGVs in Unknown Cluttered Environments: log-MPPI Control Strategy," Jul. 2022, arXiv:2203.16599 [cs]. [Online]. Available: http://arxiv.org/abs/2203.16599

[53] V. Guzov, A. Mir, T. Sattler, and G. Pons-Moll, "Human POSEitioning System (HPS): 3D Human Pose Estimation and Self-localization in Large Scenes from Body-Mounted Sensors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2021.