2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

HRI-DGDM: Dual-Graph Guided Diffusion Model for Uncertain Human Motion Modeling in HRI

Hongquan Gui, Ming Li†, *Senior Member, IEEE*

***Abstract*—Human motion in human-robot interaction (HRI) is inherently uncertain, even when performing the same task repeatedly. This variability poses a significant challenge for prediction, as models must capture a distribution of plausible futures rather than a single deterministic trajectory. Traditional graph convolutional network based models, while effective at capturing spatial temporal dependencies, are fundamentally limited by their deterministic nature and struggle to represent this inherent motion uncertainty. To address this, diffusion models have emerged as a powerful framework for modeling uncertainty. However, their direct application to HRI is hindered by two key limitations: they often prioritize motion diversity over prediction accuracy, potentially generating physically implausible results, and they fail to adequately model the complex, multi-scale spatial temporal coupling between human and robot motions. To overcome these challenges, we propose HRI-DGDM, a HRI motion prediction framework based on a dual-graph guided diffusion model. Our method introduces a dual-graph structure—comprising a structural graph for kinematic priors and a collaboration graph learned from motion dynamics—to guide the denoising process with strong structural priors. A dedicated spatial temporal denoising network (STDN) fuses multi-scale features from both graphs through adaptive fusion and hierarchical spatial temporal modeling. Furthermore, a masking-based conditioning mechanism anchors the observed history during denoising, ensuring temporal consistency and preventing drift. Experiments on HRI scenarios demonstrate that HRI-DGDM outperforms baselines in prediction accuracy.**

I. INTRODUCTION

Human-robot interaction (HRI) is pivotal for intelligent and personalized manufacturing, combining human flexibility with robotic precision [1]. In human-centered HRI, robots must proactively adapt to human intentions for safe and efficient cooperation [2]. This necessitates accurate human motion prediction, enabling robots to anticipate future movements and avoid collisions [3].

Non-contact, skeleton-based modeling has emerged as a robust approach for human motion capture in HRI, offering advantages in data efficiency and noise resistance [4]. However, human motion in real-world HRI is inherently uncertain, even when performing the same task repeatedly. This uncertainty stems from the natural variability in human behavior, where identical initial conditions can lead to significantly different motion trajectories. This poses a critical challenge for human motion prediction in HRI [5]. Deterministic models, such as those based on recurrent or graph convolutional networks, are fundamentally limited in such scenarios [6] [7]. While effective at capturing spatial temporal dependencies, their deterministic nature makes them unable to represent the full distribution of possible futures and inadequate for modeling the intrinsic uncertainty of human motion.

While diffusion models offer a promising framework for capturing motion uncertainty, their direct application to HRI is problematic [8]. First, existing diffusion models often prioritize diversity over accuracy, potentially generating physically implausible human motions. Second, they typically lack explicit mechanisms to model the complex, multi-scale spatial temporal dependencies between human and robot motions. The robot's presence is a crucial contextual cue that should condition the prediction, but most models treat human motion in isolation.

To address these limitations, we propose HRI-DGDM, a novel HRI motion prediction framework based on a dual-graph guided diffusion model. Our key insight is to integrate strong structural priors into the diffusion process to guide the denoising network towards accurate and physically plausible predictions. Specifically, HRI-DGDM introduces a dual-graph structure: a structural graph (SG) encoding the physical adjacency relationships of the human and robot, and a collaboration graph (CG) that dynamically learns functional dependencies of human and robot joints. By embedding a spatial temporal denoising network (STDN) that leverages both graphs, our model effectively captures multi-scale spatial temporal features. A masking mechanism further ensures temporal consistency by anchoring the known past during the denoising process. The main contributions of this work are summarized as follows:

(1) We propose HRI-DGDM, a novel dual-graph guided diffusion model for human-robot interaction motion prediction that explicitly models multi-scale spatial temporal dependencies.

(2) We design a STDN as the denoising network, which explicitly integrates the structural graph and collaboration graph through a multi-scale and adaptive fusion strategy. This enables the model to dynamically leverage both physical

*This work was supported in part by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. T32-707/22-N and C7076-22G), and the Innovation and Technology Fund (Project No. ITP/046/25TI).

H. Gui is with the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China (e-mail: hongquan.gui@connect.polyu.hk).

M. Li is with (a) the Department of Industrial and Systems Engineering, (b) the Research Institute for Generative AI, and (c) the Research Institute for Advanced Manufacturing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China (e-mail: ming.li@polyu.edu.hk). †Corresponding author. He is a Member of the IEEE.

kinematics and learned interaction patterns for accurate prediction.

(3) A masking-based conditioning mechanism anchors the observed history during denoising, ensuring temporal consistency and preventing drift in the known past.

II. RELATED WORKS

A. Human Motion Modeling

Human motion modeling has been a cornerstone of computer vision and robotics, with early work focusing on deterministic prediction using recurrent neural networks [9], [10] and convolutional networks [11], [12]. To capture the inherent structure of 3D skeletons, graph convolutional networks (GCNs) have been widely adopted [13], [14]. For instance, STGCN [15] and SeS-GCN [16] leverage spatial temporal graph convolutions for motion forecasting. While effective, these deterministic models generate a single trajectory and are fundamentally incapable of representing the multi-modal uncertainty inherent in human behavior, where identical initial conditions can lead to diverse future motions [17]. To address this, diffusion models have emerged as a powerful framework for modeling complex motion distributions [18], [19], [20], demonstrating superior performance in generating plausible human motion sequences [21], [22].

B. Diffusion Models for Motion Prediction

Diffusion models have rapidly become a dominant paradigm in generative modeling due to their stable training and high sample quality. In the context of motion prediction, diffusion models learn the data distribution by gradually denoising a signal corrupted with Gaussian noise [23]. Their application spans various domains, including vehicle trajectory forecasting [24] and hand motion prediction [25]. In human motion, diffusion models like HumanMAC [26] have shown remarkable success in capturing long-term dynamics. However, a key challenge for diffusion models is the trade-off between diversity and accuracy. Unconstrained diffusion processes can generate physically implausible motions, which is particularly detrimental in safety-critical applications like human-robot interaction. Recent efforts have focused on incorporating physical priors [27] or using conditional guidance [28] to improve realism. Despite these advances, existing diffusion models for human motion often treat the task in isolation and fail to condition the generation process on crucial contextual cues from the environment, such as the presence and motion of a collaborative robot.

C. Human Robot Interaction

In HRI, motion prediction is critical for enabling robots to act proactively and safely. Early HRI models extended single-agent prediction to include robot states as input, but often failed to explicitly model the bidirectional coupling between human and robot motions [29]. More sophisticated approaches have introduced shared representation spaces. For example, ECHO [30] leverages a shared latent space and attention mechanisms to forecast human motion conditioned on robot states, enabling more natural interactions. However, most existing HRI models remain deterministic [31], [32], making them ill-suited for the uncertainty of real-world collaboration. While diffusion models offer a solution for uncertainty modeling, their application to HRI is nascent. Current diffusion model based approaches [26] rarely incorporate explicit mechanisms to model the multi-scale spatial temporal dependencies of human-robot motions, nor do they leverage structural priors of the human-robot to guide the generation process. This gap motivates our work, which integrates a dual-graph structure into a diffusion framework to simultaneously capture uncertainty and enhance prediction accuracy.

III. METHODOLOGY

Human motion is inherently variable and uncertain, posing challenges for HRI. To address this, we propose the HRI-DGDM model, which embeds motion uncertainty into noise and learns the underlying human motion distribution through a denoising process. By progressively adding and then removing noise, the model captures both the temporal-spatial dependencies between human and robot motions and the statistical patterns of human behavior. Unlike prior methods, HRI-DGDM model explicitly models human-robot interdependencies, improving robustness and prediction performance for human motions.

A. HRI-DGDM Model Structure

The structure of HRI-DGDM model is shown in Fig.1. In the diffusion process, the input human-robot motions are progressively perturbed by iteratively adding Gaussian noise according to a predefined noise schedule. This Markovian noising procedure continues until the data distribution converges to an approximately standard Gaussian, thereby encapsulating the inherent uncertainty in human motion within the noise. Then, a step-by-step denoising process is applied to recover the original data distribution, wherein each denoising step is modeled by a carefully designed denoising network p_θ, whose architecture plays a critical role in ensuring prediction accuracy. To fully capture spatial temporal dependencies while avoiding over-diverse predictions, we integrate the STDN as the backbone. The STDN comprises two core components: the spatial graph convolution (SGC) module, which extracts multi-scale spatial features, and the temporal graph convolution (TGC) module, which models dynamic temporal features. By leveraging this architecture, the HRI-DGDM model effectively reduces uncertainty and generates accurate predictions of future human motion through iterative denoising.

Specifically, the observed human and robot motion are $\hat{X}_0 \in \mathbb{R}^{T_{obs} \times 3 \times J_{HR}}$, where the T_{obs} is the number of observed time steps, the J_{HR} is the number of the joints of human and robot. The outputs are the predicted future human motion $\hat{X}_0 \in \mathbb{R}^{T_{pred} \times 3 \times J_H}$, where the T_{pred} is the prediction horizon, the J_H is the number of the human joints. The inputs to the HRI-DGDM model are a complete motion sequence $X_{seq} \in \mathbb{R}^{(T_{obs}+T_{pred}) \times J_{HR} \times 3}$, formed by concatenating the observed human-robot motion $X_{HR} \in \mathbb{R}^{T_{obs} \times J_{HR} \times 3}$ and a padded future human motion $X_{padding} \in \mathbb{R}^{T_{pred} \times J_{HR} \times 3}$, the $X_{padding}$ is generated by repeating the last observed human pose across the prediction horizon — a strategy we refer to as "last-frame padding". Then, a masking mechanism is applied in the denoising process. The portion corresponding to the observed

motion ($t \leqslant T_{obs}$) is dynamically anchored by the observation, effectively "pinning" the model to the real historical motion. Meanwhile, the future portion ($t > T_{obs}$) is progressively denoised to generate a plausible and coherent future trajectory. This design ensures the model remains strongly conditioned on the observed motion, preventing drift in the known past. The padding provides a physically plausible initialization, and the mask ensures temporal consistency and accurate conditioning.

The training of the HRI-DGDM model follows the standard denoising diffusion probabilistic model framework. The model is trained to predict the Gaussian noise that is progressively added to the ground truth motion data X_0. The objective is to learn the noise prediction network STDN $\epsilon_\theta(X_t, t)$ by minimizing the mean squared error between the true noise ϵ and the predicted noise. The training loss is given by:

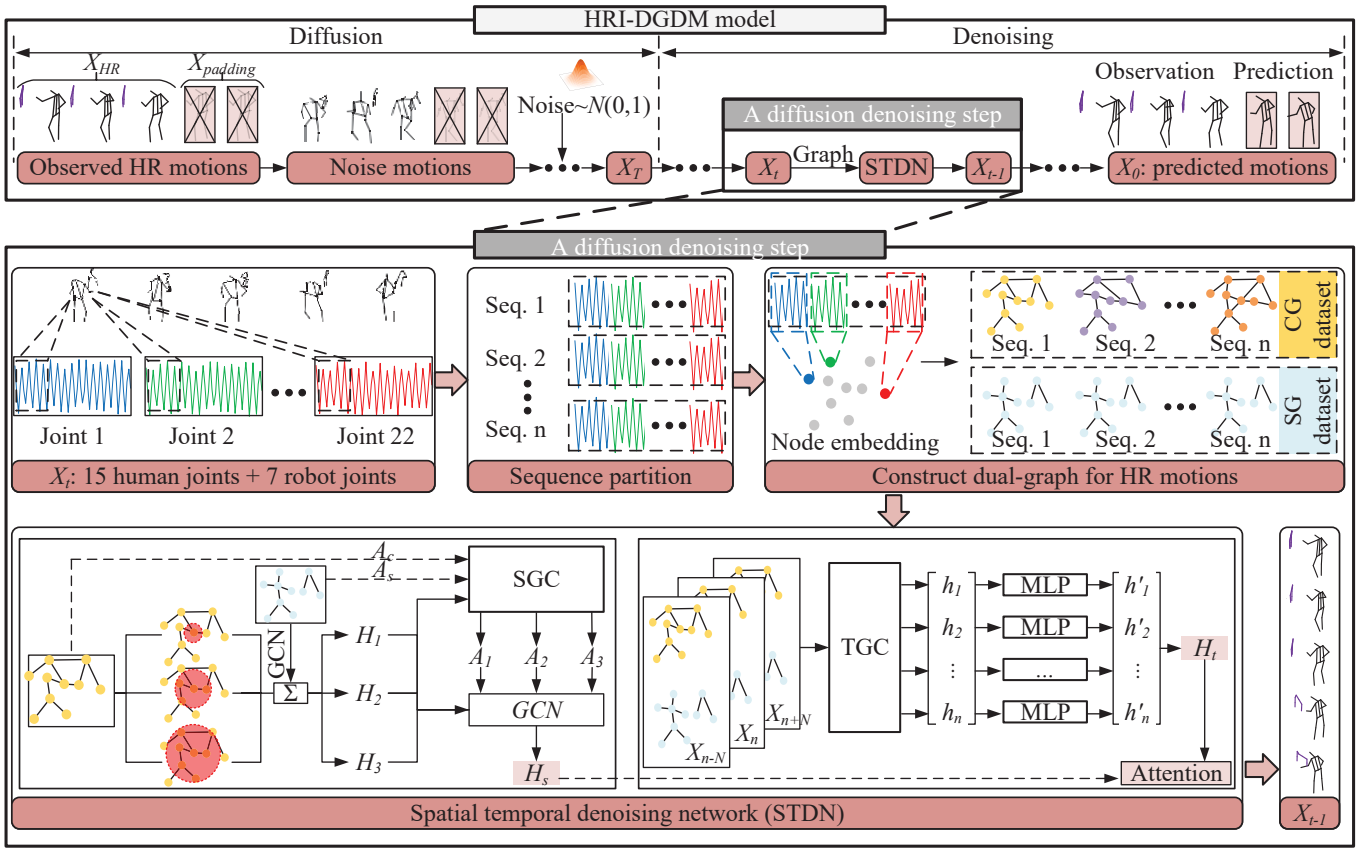

Fig. 1. Overview of the proposed HRI-DGDM model. The framework takes a complete motion sequence as input, which is formed by concatenating the observed human-robot motion and a padding future sequence. During the iterative denoising process, the STDN leverages a dual-graph structure—comprising the SG and the CG—to capture multi-scale spatial temporal features. A masking mechanism is applied at each denoising step: it preserves the observed sequence (marked as 1) and allows the model to refine the predicted sequence (marked as 0), ensuring temporal consistency and preventing drift in the known past. The final output is a refined, high-accuracy prediction of future human motion. (Seq. is short for "sequence".).

$$L = \left\| \epsilon - \epsilon_\theta \left(X_t, t \right) \right\|^2 = \left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t \right) \right\|^2 \quad (1)$$

The prediction procedure is shown in Algorithm1. The conditioning is achieved through the mask applied during the inference process, rather than being inherent to the network architecture itself. $X_{t-1} = M \odot X^o{}_{t-1} + (1 - M) \odot X^p{}_{t-1}$, where $X^o{}_{t-1}$ is the noisy observation and $X^p{}_{t-1}$ is the denoised prediction. $X^o{}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} X_{seq} + \sqrt{1 - \bar{\alpha}_{t-1}}\, z$ explicitly states that $X^o{}_{t-1}$ is derived from the observed sequence by adding controllable noise, and brings the "observed information" into the computation at every step. By using the mask M, the algorithm ensures the final output is taken from $X^o{}_{t-1}$, thereby "anchoring" the prediction to the true historical motion.

B. *Dual-graph Construction for HR Motions*

Before detailing the architecture of the STDN, we first describe the construction of the dual-graph representation (SG and CG), which serve as structural priors guiding the spatial temporal feature learning within the denoising process. Traditional diffusion models tend to prioritize diversity over accuracy in their predictions, which can lead to physically implausible forecasts in HRI scenarios. To enhance the accuracy of the diffusion model, we introduce SG and CG. The SG encodes the physical kinematic structure based on prior knowledge of keypoint connectivity in the human body and robotic arm, preserving plausible joint relationships. The collaboration graph, on the other hand, adaptively captures latent dependencies between functionally related joints by leveraging dynamic features from the motion sequences. By constructing this dual-graph representation, the denoising process is guided through the STDN, enabling more accurate motion prediction.

During the model training process, the input HR motion sequences to the denoising network are often long, making it difficult to capture local dynamic characteristics and detrimental to the model's generalization. Therefore, we employ a sliding window to partition the continuous HR motion time series X_t into n non-overlapping time windows. For each time window X_n, we treat the motion trajectory of each joint j within that window as a node, and use a learnable embedding function $\phi(\cdot)$ to map the raw joint trajectory into a high-dimensional embedding space, thereby obtaining the feature vector x_j for each node to construct the CG.

SG and CG: To effectively model the spatial temporal dependencies of HR motions, we construct two complementary graph structures: SG and CG. The SG A_s is constructed based on the joint connectivity of the human body and the robotic arm, encoding the physical adjacency relationships between joints. Nodes in the graph represent the keypoints of the HR motion, and edges represent the direct physical connections between these keypoints. This graph structure ensures the model can learn the structural constraints of human motion. The CG A_c is constructed based on the collaborative relationship between key nodes. Its adjacency matrix is calculated by the cosine similarity of the key node feature vectors:

$$A_s(i,j)=\begin{cases}1, \text{if joints } i \text{ and } j \text{ are physically connected}\\ 0, \text{otherwise}\end{cases}$$

$$A_c(i,j)=\begin{cases}\dfrac{x_i\cdot x_j}{\|x_i\|\|x_j\|}, \text{if } j \text{ is one of the most similar nodes to } i\\ 0, \text{otherwise}\end{cases} \quad (2)$$

where x_i, x_j represent the feature vector of key nodes i and j. The Top-k strategy is used to retain the most relevant k neighbors of each node. Specifically, we retain the top 10%, 20%, and 30% most relevant neighbors of each node, denoted as A_{s1}, A_{s2}, A_{s3}, respectively. This adaptive strategy ensures that the CG scales well with motion complexity.

C. Spatial Temporal Denoising Network

To ensure the denoising process is more biased towards accuracy rather than diversity, this paper proposes spatial temporal denoising network, as illustrated in Fig. 1, including SGC module and TGC module. The proposed module integrates the CG and the SG through the multi-level fusion strategy, operating at both spatial and temporal levels. This design facilitates the extraction of multi-scale features, thereby enhancing the model's capacity to comprehend HR motions.

SGC module: it captures information at different levels (local, medium-range, global) in HR motions. We construct three spatial subgraphs with k=1,2,3, denoted as A_{c1}, A_{c2}, A_{c3} and perform multi-scale graph convolution operations:

$$H_k=\mathrm{GCN}(X_n,A_s)+\mathrm{GCN}(X_n,A_{ck}),\quad k=1,2,3 \quad (3)$$

where, GCN(·) denotes a standard graph convolution operation with layer-specific learnable weights; H_k is the graph convolution output at the k-th scale; when k=1, the model aggregates information only from the top 10% most relevant spatial neighbors, allowing it to effectively capture local correlations; when k=2, the top 20% most relevant neighbors are incorporated to model medium-range dependencies and reduce interference from irrelevant nodes; when k=3, the top 30% most relevant neighbors are used to facilitate the aggregation of global spatial relationships across the pattern.

Algorithm1. Prediction for HRI-DGDM model

Input: the trained noise prediction network STDN ϵ_θ, noising steps T, the mask $M=[\underbrace{1,1,\ldots,1}_{obs},\underbrace{0,0,\ldots0}_{pred}]^\top$, observed human-robot motion data $X_{HR}\in\mathbb{R}^{T_{obs}\times3\times J_{HR}}$

Output: predicted motion data $\hat{X}_0\in\mathbb{R}^{T_{pred}\times3\times J_H}$.

$X_T=\mathcal{N}(0,I)$

$X_{seq}:=padding(X_{HR})\in\mathbb{R}^{(T_{obs}+T_{pred})\times J_{HR}\times3}$

For $t\in T,T-1,\ldots,1$ do

$z\sim\mathcal{N}(0,I)$ if $t>1$ else $z=0$

$X^o{}_{t-1}=\sqrt{\bar{\alpha}_{t-1}}X_{seq}+\sqrt{(1-\bar{\alpha}_{t-1})}z$ //condition

$X^p{}_{t-1}=\dfrac{1}{\sqrt{\alpha_t}}\left(X_t-\dfrac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_{t-1}}}\epsilon_\theta(X_t,t)\right)+\sigma_t z$

$X_{t-1}=M\odot X^o{}_{t-1}+(1-M)\odot X^p{}_{t-1}$

End

Return $\hat{X}_0$

To achieve dynamic adaptation of the graph structure, SGC module generates three score vectors P_1, P_2, P_3 from the extracted fused features, which regulate the fusion proportion between CG and SG. The W_p, b_p are learnable parameters. k=1, 2, 3. The final adjacency matrix A_{fk} for the k-th scale is computed as a weighted combination:

$$\begin{aligned}P_k&=Sigmoid\left(\mathrm{ReLU}\left(H_kW_p+b_p\right)\right)\\ A_{fk}&=P_kA_s+(1-P_k)A_{ck}\end{aligned} \quad (4)$$

The results across all scales are concatenated to produce the final structure-fused feature output:

$$H_s=GCN(H_1,A_{f1})\|\,GCN(H_2,A_{f2})\|\,GCN(H_3,A_{f3}) \quad (5)$$

TGC module: While the SCG module captures spatial dependencies, HR motions also exhibit temporal features along trajectory. To model this behavior, we introduce the TGC module that operates along the sequence of the motion data.

$$\begin{aligned}H_c^n&=\theta H_{s,l}^n+(1-\theta)H_{c,l}^n\tilde{A}_c\\ H_s^n&=\theta H_{c,l}^n+(1-\theta)H_{s,l}^n\tilde{A}_s\end{aligned} \quad (6)$$

where H_c^n, H_s^n are the features obtained from the interaction between the CG and SG along the n-th time windows, respectively; $H_{c,l}^n$, $H_{s,l}^n$ are the hidden state of CG and SG at l layer, respectively; θ is the hyperparameter, it controls the information guidance ratio between graph structures; $\tilde{A}_c$ $\tilde{A}_s$ are normalized adjacency matrices of CG and SG. Then, to

integrate information from the current and preceding time windows to filter out specific sequential features, improving the model's sequential representation capability.

$$h_c^n = f\left(\sigma\left(W_c\left[H_c^{n-1}, H_{c,l}^n\right]\right)\right)\cdot H_{c,l}^n$$
$$h_s^n = f\left(\sigma\left(W_s\left[H_s^{n-1}, H_{s,l}^n\right]\right)\right)\cdot H_{s,l}^n \quad (7)$$

where h_c^n, h_s^n are the enhanced sequence features, W_c, W_s are the learnable parameters; $f(\cdot)$, $\sigma(\cdot)$ are the Sigmoid and ReLU activation functions respectively. Combining the above two parts, the final enhanced graph interaction feature is expressed as follows:

$$H_c^n{}' = \theta H_{s,l}^n + (1-\theta) H_{c,l}^n \tilde{A}_s + h_c^n$$
$$H_s^n{}' = \theta H_{c,l}^n + (1-\theta) H_{s,l}^n \tilde{A}_g + h_s^n \quad (8)$$

Then, we concatenate the enhanced graph structures and extract sequence-fused feature representations through the multi-layer perceptron:

$$H_t = [MLP(H_c^1{}' \| H_s^1{}'), MLP(H_c^2{}' \| H_s^2{}'), \cdots, MLP(H_c^n{}' \| H_s^n{}') \quad (9)$$

The output X_{t-1} is expressed as:

$$X_{t-1} = FC\left(Attention\left(H_s, H_t\right)\right) \quad (10)$$

where attention mechanism aims to concatenate H_s and H_t to fuse the temporal and spatial features, and enables the model to dynamically focus on the most relevant graph features of the input data; the *FC* is fully connected layer, it takes the high-dimensional, fused feature representation generated by the attention mechanism and maps it back to the original data space, producing the final output X_{t-1}.

IV. EXPERIMENTS

A. Experiment Setup and Parameters Setting

1) Experiment Setup: To enable safe and efficient HRI in customized garment manufacturing, we developed the human robot collaborative cutting (HRCC) platform, as illustrated in Fig. 2. The system centers on a UR10e robotic arm equipped with a custom cutting tool and mounted on a linear sliding rail, ensuring full coverage of the fabric workspace. An assistive UR5e robot, fitted with a gripper, supports the operator by transporting tools and fabric pieces, reducing manual handling effort. The perception module integrates two RGB-D cameras and a real-time 3D human pose estimation system based on Voxelpose [39], which captures comprehensive spatial data of the operator. A digital human model is constructed using key joints, enabling dynamic tracking of human motion in the shared workspace. Robot joints are covering the base and six rotational axes. An LED interaction screen provides intuitive visual feedback, displaying task instructions and safety alerts. The proposed HRI-DGDM predicted the human motion to detect collision. When a potential collision is detected (e.g., during time intervals *T*1 to *T*3), the robot automatically reduces speed until pauses operation.

2) Parameters Setting: Our experiments are conducted on a Linux system with an Intel Xeon W-2295 CPU and 128GB DDR4 memory. We implement the HRI-DGDM model in Python using the PyTorch framework. The batch size is set to 32, and the hidden dimension to 256. The initial learning rate is 0.0001, optimized using AdamW. The diffusion process is configured with 500 noise steps and a time step of 50. The model uses a dropout rate of 0.2 to mitigate overfitting. Training runs for 80 epochs with an exponential decay learning rate scheduler.

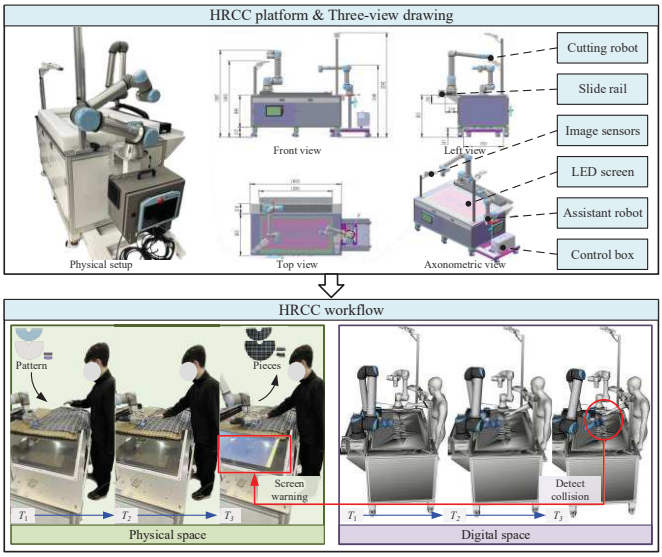

Fig. 2. The experimental setup in the HRCC scenario.

3) Dataset: To comprehensively evaluate the effectiveness of our human motion modeling approach, we conduct experiments on two datasets: the public HRI dataset CHICO and our privately collected HRCC dataset. First, on the HRCC dataset, we perform 10 experimental trials in a fabric-cutting task, recording the coordinates of both human and robot joint positions. One representative trial is illustrated in Fig. 3. The results show that a safe distance is maintained between the human and robot along the X-axis throughout the interaction. While the human is arranging the fabric, the robot executes three consecutive cutting motions. Notably, the robot demonstrates significantly higher motion repeatability compared to the human. The robot's joint trajectories exhibit consistent repetition accuracy, whereas human joint movements—especially in upper-limb joints such as the right wrist (X_{11})—show substantial variability across repetitions. This low repeatability may stem from inherent human behavioral uncertainty. This variability underscores the necessity of modeling human motion uncertainty explicitly. Furthermore, the spatial temporal dependencies between human and robot motions highlight the importance of capturing dynamic patterns. Incorporating such spatial temporal features into motion models is essential for accurate human behavior understanding in collaborative settings.

Due to the scarcity of public datasets in HRC, we also validate our approach on the CHICO dataset [16]. While prior studies on CHICO focus on predicting single, repetitive actions, we extend its use by integrating multiple action types into a unified framework for multi-action mixed HRC scenarios. This enables us to evaluate our model's ability to handle complex, real-world interactions involving task variation, coordination, and intermittent collaboration. By validating on both the HRCC and CHICO datasets, we ensure

robust assessment of our modeling approach across different interaction patterns, environmental setups, and levels of human motion uncertainty.

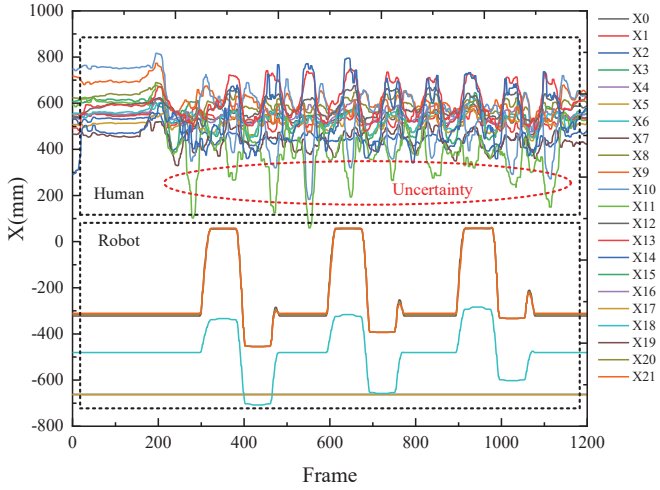

Fig. 3. Visualization of the uncertainty in human-robot motion data. The sequences depict the trajectories of the human operator (joints X_0–X_{14}) and the robot (joints X_{15}–X_{21}) across multiple repetitions of the same task. The significant variation in the human trajectories, despite identical initial conditions and task goals, highlights the inherent uncertainty of human motion.

B. Quantitative Evaluation

To evaluate prediction performance, we compare our model against several representative baselines: MRS-GCN, STS-GCN, SeS-GCN, ECHO, and HumanMac, with our proposed HRI-DGDM as the focus of comparison. We adopt the mean per joint position error (MPJPE) as the primary evaluation metric, which measures the average Euclidean distance between predicted and ground-truth joint positions, providing a reliable assessment of spatial accuracy in human motion prediction. The comparison results are shown in Tab. 1.

TABLE I: Quantitative evaluation of the short-term (400ms observation, 400ms prediction) and long-term (400ms observation, 1000ms prediction) motion prediction in the CHICO dataset and HRCC dataset reported in MPJPE. Here, bold indicates the best result.

	CHICO		**HRCC**	
Milliseconds	**400**	**1000**	**400**	**1000**
MRS-GCN[32]	54.7	91.4	48.8	85.6
STS-GCN[31]	54.1	87.8	48.3	82.1
SeS-GCN[16]	48.9	85.7	43.6	80.3
ECHO[30]	47.3	81.7	41.9	76.5
HumanMac[26]	46.9	81.2	41.3	75.8
HRI-DGDM	**45.7**	**79.8**	**39.6**	**74.4**

The results indicate that while GCN-based models such as MRS-GCN and STS-GCN are capable of modeling spatial temporal relationships, they may produce over-smoothed motions, leading to a loss of fine-grained dynamics and naturalness. Moreover, these models lack effective mechanisms for fusing features across human and robot motion data. SeS-GCN achieves competitive performance, yet its limitations highlight a common challenge among existing approaches. The performance gap between SeS-GCN and ECHO in long-term prediction reveals the inherent difficulty of traditional models in capturing long-range temporal dependencies. The ECHO model improves upon this by leveraging self-attention and cross-attention mechanisms within a shared human-robot representation space, enabling more effective learning of dynamic coupling between human and robot motions. However, it still struggles to adequately handle the intrinsic uncertainty present in human motion. Moreover, it did not model the multi-scale spatial temporal dependencies of human-robot motions. HumanMac, built upon a diffusion modeling framework, excels at capturing the uncertainty and rich spatial temporal patterns of human motion. Nevertheless, as it was not originally designed for HRI, it fails to fully account for the multi-scale spatial temporal dependencies between human and robot motions. This results in predictions that prioritize diversity over accuracy.

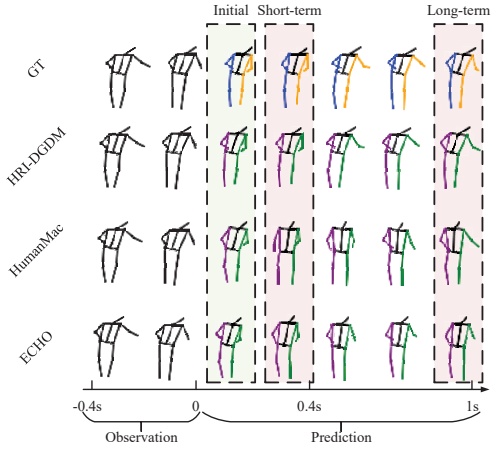

Fig. 4. Qualitative comparison of human motion prediction at the initial (200ms), short-term (400ms), and long-term (1000ms) horizons. The observed history is shown in black, the ground truth future is in blue and yellow, the prediction motion is in green and purple. HRI-DGDM consistently predicts accurate and physically plausible predictions across all time scales. At the initial stage, all models perform well, closely following the ground truth. However, as the prediction horizon extends, a clear performance divergence emerges: HumanMAC, while generating diverse motions, tends to produce trajectories that deviate from the ground truth and exhibit implausible poses. ECHO struggles to capture the full spectrum of motion dynamics, resulting in inaccurate long-term predictions. In contrast, HRI-DGDM maintains high stability.

In contrast, our proposed HRI-DGDM model operates within a diffusion framework while incorporating a carefully designed STDN. By explicitly modeling both motion uncertainty and multi-scale spatial temporal dependencies in human-robot interactions, HRI-DGDM achieves superior performance. The dual-graph structure—comprising a SG based on physical kinematics and a CG learned from motion dynamics—provides strong structural priors that guide the denoising process toward physically plausible and accurate predictions. Furthermore, the integration of spatial and temporal graph convolutions within the STDN enables hierarchical feature extraction, enhancing the model's ability to capture local, medium-range, and global spatial temporal patterns. Through iterative denoising conditioned on observed motion via a masking mechanism, HRI-DGDM maintains temporal consistency and avoids drift in the known past, while generating diverse yet realistic future motions. Therefore, HRI-DGDM achieves the lowest MPJPE across both short- and long-term prediction horizons. The qualitative results, as shown in Fig. 4, further illustrate the superior performance of our method.

C. Ablation Study

To evaluate the effectiveness of the key components of the HRI-DGDM model, we conducted an ablation study focusing on three critical elements: the diffusion framework, the SGC module, and the TGC module. The results of this study are presented in Tab. 2. The experimental results reveal several important insights. First, the performance of HRI-DGDM model without TGC is slightly lower than that of the full HRI-DGDM, indicating the significance of the TGC. This module plays a role in capturing temporal human-robot motion features, preventing the gradual weakening of motion features over time. Second, the prediction accuracy of HRI-DGDM model without SGC, is significantly lower than that of HRI-DGDM model without TGC, demonstrating that the SGC module is more critical than the TGC module. The SGC is designed to comprehensively fuse multi-scale spatial features, enabling the model to better capture the comprehensive interactions between human and robot motions. Finally, the prediction accuracy of HRI-DGDM without diffusion framework is the lowest among all variants, highlighting the diffusion framework as the most important component. The diffusion framework excels at capturing the uncertainty of human motions, providing a robust foundation for accurate predictions. By gradually adding and then reversing noise, it effectively models uncertainty as part of the stochastic process, allowing the model to recover realistic motion patterns through iterative denoising.

TABLE II: Ablation study of HRI-DGDM model with different components removed for motion prediction in CHICO and HRCC datasets

	CHICO		**HRCC**	
Milliseconds	**400**	**1000**	**400**	**1000**
w/o diffusion framework	48.0	81.0	40.9	75.5
w/o SGC module	47.2	80.3	40.1	75.1
w/o TGC module	46.1	80.1	39.7	74.8
HRI-DGDM	**45.7**	**79.8**	**39.6**	**74.4**

V. Conclusion

In this work, we propose HRI-DGDM, a novel HRI motion prediction framework that effectively captures both the uncertainty inherent in human motion and the complex multi-scale spatial temporal dependencies of human-robot motions. By embedding a dual-graph guided STDN into a diffusion-based generative process, our model leverages a structural graph to preserve kinematic constraints and a dynamic collaboration graph to learn functional inter-joint relationships, enabling robust and physically plausible motion forecasting. The SGC and TGC modules allow HRI-DGDM to model local, medium-range, and global spatial temporal patterns, while the masking-based conditioning strategy ensures temporal consistency by anchoring the observed history during denoising. The experiments demonstrate that HRI-DGDM outperforms existing state-of-the-art methods in prediction accuracy. This work highlights the importance of integrating structural priors and interaction-aware feature learning into generative models for reliable and natural human-robot motion prediction.

While the collaboration graph captures dynamic human-robot dependencies through data-driven similarity, its construction relies on pre-segmented motion windows and a fixed Top-k sparsity assumption, which may not fully adapt to abrupt changes in interaction dynamics or varying motion frequencies across different tasks. We plan to investigate adaptive graph learning mechanisms to enable finer-grained, event-triggered evolution of the collaboration structure, and to extend HRI-DGDM into real-time, closed-loop interaction systems with online trajectory refinement

References

[1] Wang, L., Gao, R., Váncza, J., Krüger, J., Wang, X. V., Makris, S., & Chryssolouris, G. (2019). Symbiotic human-robot collaborative assembly. CIRP annals, 68(2), 701-726. https://doi.org/10.1016/j.cirp.2019.05.002.

[2] Inkulu, A. K., & Bahubalendruni, M. R. (2024). Human-robot collaborative task planning for assembly system productivity enhancement. Robotic Intelligence and Automation. https://doi.org/10.1108/RIA-05-2023-0067.

[3] Li, S., Wang, R., Zheng, P., & Wang, L. (2021). Towards proactive human–robot collaboration: A foreseeable cognitive manufacturing paradigm. Journal of Manufacturing Systems, 60, 547-552. https://doi.org/10.1016/j.jmsy.2021.07.017.

[4] Zhang, R., Li, J., Zheng, P., Lu, Y., Bao, J., & Sun, X. (2022). A fusion-based spiking neural network approach for predicting collaboration request in human-robot collaboration. Robotics and Computer-Integrated Manufacturing, 78, 102383. https://doi.org/10.1016/j.rcim.2022.102383.

[5] Zhang, R., Lv, J., Li, J., Bao, J., Zheng, P., & Peng, T. (2022). A graph-based reinforcement learning-enabled approach for adaptive human-robot collaborative assembly operations. Journal of Manufacturing Systems, 63, 491-503. https://doi.org/10.1016/j.jmsy.2022.05.006.

[6] Schydlo, P., Rakovic, M., Jamone, L., & Santos-Victor, J. (2018, May). Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 5909-5914). IEEE. https://doi.org/10.1109/ICRA.2018.8460924.

[7] Wang, P., Liu, H., Wang, L., & Gao, R. X. (2018). Deep learning-based human motion recognition for predictive context-aware human-robot collaboration. CIRP annals, 67(1), 17-20. https://doi.org/10.1016/j.cirp.2018.04.066.

[8] Ahn, H., Mascaro, E. V., & Lee, D. (2023, May). Can we use diffusion probabilistic models for 3d motion prediction? In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 9837-9843). IEEE. https://doi.org/10.1109/ICRA48891.2023.10160722.

[9] Liu, J., Wang, G., Duan, L. Y., Abdiyeva, K., & Kot, A. C. (2017). Skeleton-based human action recognition with global context-aware attention LSTM networks. IEEE Transactions on Image Processing, 27(4), 1586-1599. https://doi.org/10.1109/TIP.2017.2785279.

[10] Zhang, J., Liu, H., Chang, Q., Wang, L., & Gao, R. X. (2020). Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly. CIRP annals, 69(1), 9-12. https://doi.org/10.1016/j.cirp.2020.04.077.

[11] Ke, Q., Bennamoun, M., An, S., Sohel, F., & Boussaid, F. (2018). Learning clip representations for skeleton-based 3d action recognition. IEEE Transactions on Image Processing, 27(6), 2842-2855. https://doi.org/10.1109/TIP.2018.2812099.

[12] Zhang, J., Wang, P., & Gao, R. X. (2021). Hybrid machine learning for human action recognition and prediction in assembly. Robotics and Computer-Integrated Manufacturing, 72, 102184. https://doi.org/10.1016/j.rcim.2021.102184.

[13] Terreran, M., Barcellona, L., & Ghidoni, S. (2023). A general skeleton-based action and gesture recognition framework for human–robot collaboration. Robotics and Autonomous Systems, 170, 104523. https://doi.org/10.1016/j.robot.2023.104523.

[14] Yan, S., Xiong, Y., & Lin, D. (2018, April). Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1). https://doi.org/10.1609/aaai.v32i1.12328.

[15] Zhang, Y., Ding, K., Hui, J., Lv, J., Zhou, X., & Zheng, P. (2022). Human-object integrated assembly intention recognition for context-aware human-robot collaborative assembly. Advanced Engineering Informatics, 54, 101792. https://doi.org/10.1016/j.aei.2022.101792.

[16] Sampieri, A., di Melendugno, G. M. D. A., Avogaro, A., Cunico, F., Setti, F., Skenderi, G., ... & Galasso, F. (2022, October). Pose forecasting in industrial human-robot collaboration. In European Conference on Computer Vision (pp. 51-69). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19839-7_4.

[17] Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., & Liu, Z. (2024). Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2024.3355414.

[18] Dabral, R., Mughal, M. H., Golyanik, V., & Theobalt, C. (2023). Mofusion: A framework for denoising-diffusion-based motion synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9760-9770). https://doi.org/10.1109/CVPR52729.2023.00941.

[19] Liang, H., Zhang, W., Li, W., Yu, J., & Xu, L. (2024). Intergen: Diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision, 1-21. https://doi.org/10.1007/s11263-024-02042-6

[20] Yuan, Y., Song, J., Iqbal, U., Vahdat, A., & Kautz, J. (2023). Physdiff: Physics-guided human motion diffusion model. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 16010-16021). https://doi.org/10.48550/arXiv.2212.02500.

[21] Barquero, G., Escalera, S., & Palmero, C. (2023). Belfusion: Latent diffusion for behavior-driven human motion prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2317-2327). https://doi.org/10.48550/arXiv.2211.14304.

[22] Liang, H., Zhang, W., Li, W., Yu, J., & Xu, L. (2024). Intergen: Diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision, 132(9), 3463-3483. https://doi.org/10.1007/s11263-024-02042-6.

[23] Wei, D., Sun, H., Li, B., Lu, J., Li, W., Sun, X., & Hu, S. (2023, June). Human joint kinematics diffusion-refinement for stochastic motion prediction. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 5, pp. 6110-6118). https://doi.org/10.1609/aaai.v37i5.25754.

[24] Li, Z., Liang, H., Wang, H., Zheng, X., Wang, J., & Zhou, P. (2023). A multi-modal vehicle trajectory prediction framework via conditional diffusion model: A coarse-to-fine approach. Knowledge-Based Systems, 280, 110990. https://doi.org/10.1016/j.knosys.2023.110990.

[25] Tang, B., Zhang, K., Luo, W., Liu, W., & Li, H. (2024, September). Prompting future driven diffusion model for hand motion prediction. In European Conference on Computer Vision (pp. 169-186). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72667-5_10.

[26] Chen, L. H., Zhang, J., Li, Y., Pang, Y., Xia, X., & Liu, T. (2023). Humanmac: Masked motion completion for human motion prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9544-9555). https://doi.org/10.48550/arXiv.2302.03665.

[27] Karunratanakul, K., Preechakul, K., Suwajanakorn, S., & Tang, S. (2023). Guided motion diffusion for controllable human motion synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2151-2162). https://doi.org/10.48550/arXiv.2305.12577.

[28] Yu, H., Hou, Y., Pei, W., Ong, Y. S., & Zhang, Q. (2024). Divdiff: A conditional diffusion model for diverse human motion prediction. IEEE Transactions on Multimedia. https://doi.org/10.48550/arXiv.2409.00014.

[29] Lee, D., Ott, C., & Nakamura, Y. (2010). Mimetic communication model with compliant physical contact in human—humanoid interaction. The International Journal of Robotics Research, 29(13), 1684-1704. https://doi.org/10.1177/027836491036416.

[30] Mascaro, E. V., Yan, Y., & Lee, D. (2024, May). Robot interaction behavior generation based on social motion forecasting for human-robot interaction. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 17264-17271). IEEE. https://doi.org/10.48550/arXiv.2402.04768.

[31] Sofianos, T., Sampieri, A., Franco, L., & Galasso, F. (2021). Space-time-separable graph convolutional network for pose forecasting. In 2021 IEEE. In CVF International conference on computer vision (ICCV) (Vol. 1, p. 4). https://doi.org/10.48550/arXiv.2110.04573

[32] Dang, L., Nie, Y., Long, C., Zhang, Q., & Li, G. (2021). Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 11467-11476). https://doi.org/10.48550/arXiv.2108.07152.