Multi-agent Long-term 3D Human Pose Forecasting via Interaction-aware Trajectory Conditioning

Jaewoo Jeong*, Daehee Park*, and Kuk-Jin Yoon
KAIST
{jeong207,bag2824,kjyoon}@kaist.ac.kr

Abstract

*Human pose forecasting garners attention for its diverse applications. However, challenges in modeling the multi-modal nature of human motion and intricate interactions among agents persist, particularly with longer timescales and more agents. In this paper, we propose an interaction-aware trajectory-conditioned long-term multi-agent human pose forecasting model, utilizing a coarse-to-fine prediction approach: multi-modal global trajectories are initially forecasted, followed by respective local pose forecasts conditioned on each mode. In doing so, our **T**rajectory**2P**ose model introduces a graph-based agent-wise interaction module for a reciprocal forecast of local motion-conditioned global trajectory and trajectory-conditioned local pose. Our model effectively handles the multi-modality of human motion and the complexity of long-term multi-agent interactions, improving performance in complex environments. Furthermore, we address the lack of long-term (6s+) multi-agent (5+) datasets by constructing a new dataset from real-world images and 2D annotations, enabling a comprehensive evaluation of our proposed model. State-of-the-art prediction performance on both complex and simpler datasets confirms the generalized effectiveness of our method. The code is available at* https://github.com/Jaewoo97/T2P.

1. Introduction

Human pose forecasting aims to predict future human motion based on observed past motion [20, 31, 32, 35, 37, 53, 86]. Humans instinctively perform such tasks, allowing them to naturally navigate in crowded areas or identify and circumvent potential dangers. For this reason, human pose forecasting plays an important role in various computer vision tasks [21, 23, 27, 54, 85, 91]. Indeed, recent years have seen a proliferation of work on multi-agent motion forecasting which aim towards modeling complex multi-agent interaction [20, 47, 53, 71, 74].

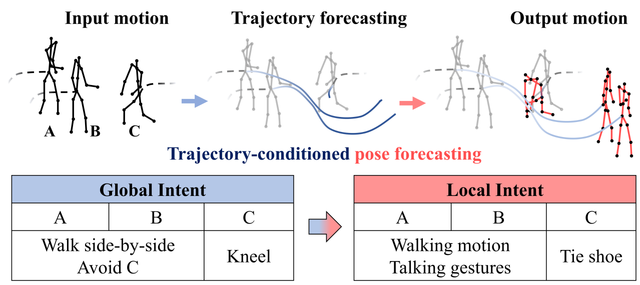

Figure 1. Human motion is goal-directed and influenced by other entities. Therefore, global intention contains hints for local intention, allowing us to infer local pose from global trajectories. Our method first forecasts global trajectories, upon which local poses are conditioned for subsequent forecasts. Pose and trajectory-wise inter-agent interactions are considered for both predictions.

Although various methods have been proposed, they share two major limitations. The first is a limitation on long-term predictions, as previous studies predicted up to 3 seconds at most [4, 47, 74, 75]. However, a sufficiently long forecast horizon is essential to fully leverage human pose forecasting for diverse downstream tasks in the scope of identifying potential danger or understanding human behavior. The second is that multi-person interactions are not proficiently learned. Existing methods consider the joints of multiple people all at once as objects of interaction [47, 65, 74], resulting in an excessive complexity with respect to the number of joints. Due to such inefficient modeling, these approaches are found to be incompetent in long-term (3s+) multi-agent (6+) settings, limiting their practicality on complex real-world environments.

Moreover, these challenges are also due to the limitations of datasets. Existing pose forecasting datasets have limited sequence length (∼3s) and number of agents (∼2). Therefore, previous works [47, 69, 75] have randomly blended disparate datasets to model multi-agent interaction with up to 10 agents. Yet, such naively merged data lacks authentic interaction as agents from different scenes remain uninfluenced. As such, there was no opportunity to develop and

*Denotes equal contribution.

evaluate a model in a long-term multi-agent environment.

To this end, we present a solution from both model and dataset perspectives to tackle long-term multi-agent human pose forecasting. First, from a model perspective, we propose an interaction-aware trajectory-conditioned pose forecasting method. We point out that the limitations of existing methods on long-term multi-agent environments lead to poor performance in handling the multi-modal nature of human motion and correspondingly complex interactions. To improve upon handling multi-modality in these complex settings, we use a coarse-to-fine approach to enjoy effective interaction modeling by propagating agent-wise coarse representations. Agent-wise pose and trajectory embeddings are obtained in their respective local coordinates, followed by a holistic interaction modeling via our proposed Traj-pose module. Interaction-aware forecasts are then made by initial **coarse** global hip joint trajectory forecast followed by **fine** local pose forecasts in its hip joint coordinates, conditioned on the global trajectory as shown in Fig. 1. As discovered in previous research [2, 54], learning an agent-wise global intention as coarse trajectories is less challenging than predicting every joint-wise motion. We leverage these hints from global trajectories, which are further conditioned towards forecasting local motion that embodies the interaction-aware spatio-temporal context.

From a dataset perspective, we parsed a novel real-world dataset for long-term multi-agent human pose forecasting. We utilize JRDB dataset [66] which consists of multi-view video and collected in various environments. Since 3D pose annotations are not provided in the original JRDB, we extracted sequences of 3D human pose from visible agents in omnidirectional images using the latest algorithm for 3D pose extraction from image [59]. We then ensure the reliability of 3D pose information by filtering and adjusting the extracted 3D poses based on 2D pose and 3D bounding box annotations. As a result, we construct a real-world 3D human pose forecasting dataset, JRDB-GlobMultiPose (JRDB-GMP), where up to 24 agents exist for up to 5 seconds. The proposed pose forecasting model is validated on both previous datasets and newly created JRDB-GMP dataset. Our method shows state-of-the-art forecasting performance in both global and local accuracy metrics, not only on JRDB-GMP but also on all previous datasets. Therefore, our contributions are as follows:

- We propose a interaction-aware trajectory-conditioned pose forecasting method (*T2P*) for long-term multi-agent 3D human pose forecasting.

- We propose a long-term, multi-agent real-world 3D human pose forecasting dataset which contains up to 24 persons and forecasts up to 5 seconds.

- We validate our *T2P* model on both previous datasets and our new JRDB-GMP dataset. Our method achieves state-of-the-art forecasting performance on all datasets.

2. Related works

2.1. Human pose forecasting

Human Pose Forecasting involves predicting a future pose sequence with temporal length of a prediction horizon, given a historical pose sequence [5, 8, 11, 16, 36, 45, 49, 52, 68, 71]. In the early stage, methods were developed to forecast single person motion within a short timeframe ($\sim$ 1s) [9, 34, 58, 73]. However, to improve applicability on diverse downstream computer vision tasks, forecasts are to be made on multi-person poses [1, 2, 63] for longer prediction horizons [6, 62]. Forecasting future inherently involves a stochastic nature, and handling such multi-modality has been attempted by forecasting multiple future poses of a single agent [4]. However, comparatively marginal efforts have been employed in the more complex long-term multi-agent scenes [75]. Such absence is mostly due to the lack of a proper dataset. The commonly used evaluation datasets are CMU-Mocap [13], 3DPW [67], UMPM [64], MuPoTS-3D [39], all of which contain 2 agents at most in a given scene and have short prediction horizons within 3 seconds. Most recent research arbitrarily combines individual scenes to create datasets with more than three individuals [47, 69, 75]. However, such a synthetic approach does not account for authentic agent interactions.

2.2. Trajectory prediction

Trajectory prediction involves predicting the future path of an object given its past trajectory [3, 10, 25, 38, 40, 41, 51, 76, 77, 89, 90]. Unlike human pose forecasting which aims to predict every joint position, trajectory prediction regards each agent as a point mass, typically the center of mass or center point of a detected bounding box. Research in trajectory forecasting is interested in not only vehicles but also many types of agents including humans, cyclists, and so on [29, 43, 72, 84]. One substantial direction in research within this field is the Goal-conditioned prediction approach [18, 33, 82]. Goal-conditioned prediction approach first predicts the final destination within the prediction horizon with multiple goal proposals [28, 70]. Then, a thorough future path is conditioned on each mode of the multi-modal proposals. Compared to directly predicting full trajectories, the goal-conditioned approach follows a coarse-to-fine prediction and is effective in learning highly stochastic multi-modality of complex scenes [19, 42, 88].

2.3. Human pose estimation from image

Human pose estimation is the task of inferring the pose of a person from an image or a video [22, 24, 26, 30, 57, 60, 78, 80, 81]. Initial deep learning-based methods first utilized convolutional neural networks to estimate 2D and 3D poses from single or multiple images [12, 14, 61, 79, 83]. Recent approaches engage in more challenging tasks such as

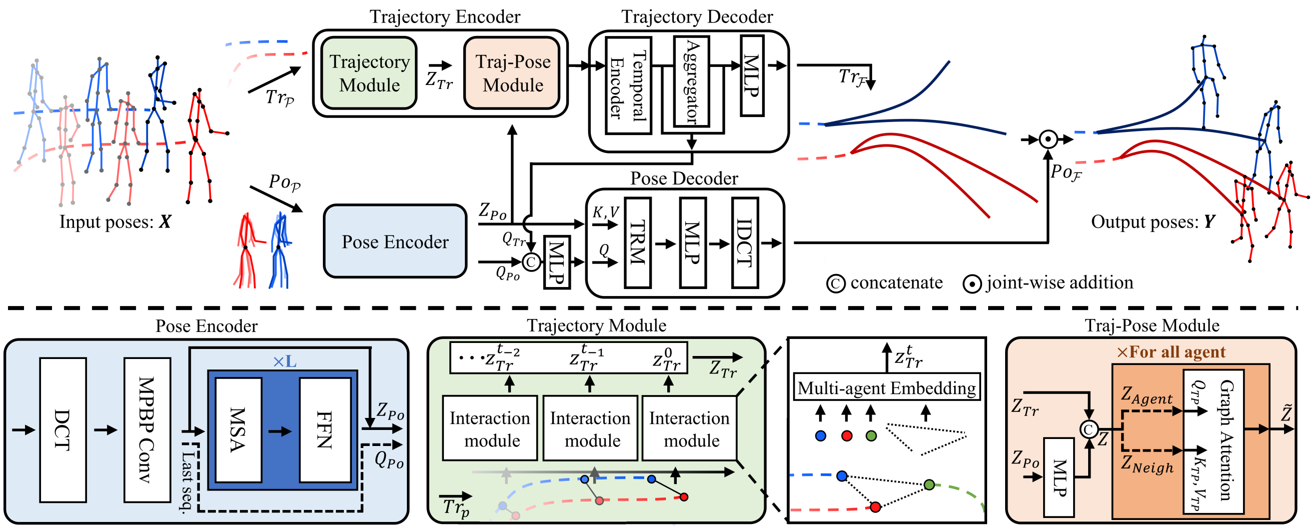

Figure 2. Illustration of our T2P framework. We decompose global motion into global trajectory and local pose. Multi-modal global trajectory proposals are predicted from past global trajectory and local pose embeddings. Then, future local poses are conditioned and forecasted on each trajectory proposal to compose the final human pose prediction. Predicted local poses are added to their mode-specific global trajectories in a joint-wise manner, obtaining the global human poses as the final output.

estimating 3D poses from monocular videos [7, 46, 48, 50, 55, 56] using self-supervised learning and generative methods [15, 17]. Most recent methods estimate multi-person poses in a crowded environment with considerable occlusions [44, 87]. We account for the aforementioned need of a complex dataset by extracting 3D pose from images using these methods. Specifically, we use a monocular 3D pose estimation method *BEV* [59] to construct a 3D human motion forecasting dataset with long-term multi-agent characteristics from real-world image sequences. *BEV* robustly estimates human pose in a scale-ambiguous and crowded environment, reliably extracting 3D poses from the omnidirectional image sequences of JRDB dataset [66].

3. Method

3.1. Problem definition

Multi-agent human pose forecasting aims to learn a mapping function between the observed 3D pose of N_A agents composed of J joints, $\mathbf{X} : \left\{\mathbf{x}_t^{n,j}\right\}_{-T_p:0}^{N_A,J}$, and future pose $\mathbf{Y} : F \times \left\{\mathbf{x}_t^{n,j}\right\}_{0:T_f}^{N_A,J}$ in global coordinates where F denotes the number of modes. Here, T_p and T_f are history length and prediction horizon while $\mathbf{x}_t^{n,j} = (x_t^{n,j}, y_t^{n,j}, z_t^{n,j})$ is 3D global coordinate of joint j of agent n at time t. While the global position of joint is represented in $\mathbf{x}$, we additionally define local position $\mathbf{p}$. The local position is defined in local coordinate of each agent, calculated by subtracting the global position of the hip joint of each agent. Therefore, local position of joint is defined as $\mathbf{p}^{n,j} = \mathbf{x}^{n,j} - \mathbf{x}^{n,\text{hip}}$. We define the trajectory of global hip joint position as global trajectory, $Tr : \left\{\mathbf{x}^{n,\text{hip}}\right\}^{N_A}$. We also define local pose as local position of all joints, $Po : \left\{\mathbf{p}^{n,\text{j}}\right\}^{N_A,J}$. We denote past and future timesteps of global trajectory and local motion as $Tr_{\mathcal{P}}, Tr_{\mathcal{F}} \in Tr$ and $Po_{\mathcal{P}}, Po_{\mathcal{F}} \in Po$, where $\mathcal{P}$ and $\mathcal{F}$ respectively denotes past and future.

3.2. Overall framework

We disentangle the overall human motion into global trajectories and local poses, as depicted in top left of Fig. 2. Following a coarse-to-fine strategy, multiple global trajectories are first forecasted to model the coarse modes of global intentions. Based on these forecasts, local pose predictions are conditioned on each mode to jointly constitute a thorough motion. In doing so, our model is widely divided into two portions: Trajectory predictor consists of trajectory encoder and decoder and pose predictor consists of pose encoder and decoder. Both predictors engage in the reciprocal exchange of both trajectory and pose information, facilitating the inference of cues between global and local motion. The detailed methods of each stage are described below:

3.3. Model structure

3.3.1 Pose encoder

Unlike the holistic approach of previous works that encode and decode all agents' joint motions in global coordinates, our pose encoder encodes the pose dynamics in local coordinates. In addition, our pose encoder only considers intra-agent joint interaction. As a result, the encoded pose embedding represents agent-specific local motion, containing insights on global intent. We follow our baseline [47] and construct the encoder with Multi-Person Body-Part (MPBP) module and transformer networks. As depicted in Fig. 2, body part sequences are constructed in frequency domain, followed by intra-agent attention-based encoding of the body parts to acquire pose embedding Z_{Po}.

3.3.2 Trajectory module

Trajectory module aims to extract embeddings from the agents' past global trajectory. Using an encoder structure from [88], multi-agent interaction-based trajectory embedding Z_{Tr} is extracted which contains rudimentary insight on global intent. Interaction between agent trajectories is represented based on the reference agent i's global trajectory segment vector $\mathbf{v}_t^i = \mathbf{x}_t^{i,\text{hip}} - \mathbf{x}_{t-1}^{i,\text{hip}}$. For rotational invariance, neighbor actor j's vector is normalized by the reference vector's orientation at latest timestep t=0. Separate MLP layers then compute the reference agent and neighboring agent embeddings $z_{Tr_i}^t, z_{Tr_j}^t$ as follows:

$$\begin{aligned} z_{Tr_i}^t &= \phi_{ref}(R_i^T \mathbf{v}_t^i) \\ z_{Tr_j}^t &= \phi_{nbr}([R_i^T(\mathbf{v}_t^j), R_i^T(\mathbf{v}_t^i)]) \end{aligned} \tag{1}$$

where ϕ_{ref} and ϕ_{nbr} are different MLP blocks, $R_i \in \mathbb{R}^{3\times3}$ is the rotation matrix of agent j against agent i, $[\cdot,\cdot]$ is concatenation. The resulting agent-specific reference and neighbor embeddings constitute trajectory embedding z_{Tr}^t.

3.3.3 Traj-pose module

Human maneuver contains various dynamic activities characterized by the agent's multi-modal intents. Auxiliary human motion such as arm gesture, rotational orientation of upper body and head implies the agent's intent in global motion. In that sense, harvesting meaningful insights from past local joint motion helps proficient modeling of coarse multi-modality as future trajectory proposals. Therefore, we propose Traj-Pose Module that fuses agent-wise embeddings of both trajectory and pose to fully utilize these information in modeling global intentions.

First, MLP is used to match the temporal domain of pose embedding Z_{Po} to that of Z_{Tr}, after which both are concatenated as agent-wise traj-pose embedding Z.

$$Z = [Z_{Tr}, \phi_{MLP}(Z_{Po})] \tag{2}$$

The resulting Z is comprised of agent and timestep-respective trajectory and pose embeddings: $z_i^t, z_j^t \in z^t \in Z$ Then, $\widetilde{Z}$ is acquired from the graph attention with an agent-wise update where each agent embedding z_i^t and its neighbor embedding z_j^t are used as query and key/value. Similar to trajectory interaction encoder of HiVT [88], the graph attention operation is operated as follows:

$$\begin{aligned} \alpha_i^t &= \text{softmax}(\frac{q_i^{t\top}}{\sqrt{d_k}} \cdot [\{k_j^t\}_{j\in N_i}]), \\ m_i^t &= \sum_{j\in N_i} \alpha_i^t v_j^t, \\ g_i^t &= \text{sigmoid}(W^{\text{gate}}[z_i^t, m_i^t]), \\ \widetilde{z}_i^t &= g_i^t \odot W^{\text{self}} z_i^t + (1-g_i^t) \odot m_i^t \end{aligned} \tag{3}$$

where N_i is a set of agent i's neighbors, W^{gate} and W^{self} are learnable matrices, and $\odot$ is element-wise product.

3.3.4 Trajectory decoder

Trajectory is forecasted from the output of trajectory encoder which encodes both past global trajectory and local pose information. Since its graph operation is operated by each timestep, a temporal encoder is used as a temporal encoder to integrate $\widetilde{Z}$ in the temporal dimension. A multi-head self-attention temporal encoder is used as the temporal encoder. Aggregator then takes into account variations in local coordinate frames to accurately represent geometric relationships within the global coordinate system via a graph operation. MLP is subsequently applied to span embedding F times for multi-modal prediction, which is residually added to the $\times F$ repeated embedding before the aggregator. Finally, another MLP is used to extract multi-modal future global trajectory proposals of hip joint $Tr_{\mathcal{F}} \in \mathbb{R}^{F\times T_f\times 3}$. The multi-modal embedding is also passed onto the Pose Decoder to forecast local poses.

3.3.5 Pose decoder

Future human pose depends on past human poses and global intention. The pose decoder is designed to consider these factors while generating local poses via mode-specific trajectory conditioning. A transformer (TRM) decoder is used to decode local motions, where pose embedding Z_{Po} is used as key/value and concatenation of trajectory and pose query.

$$Q = \phi_{MLP}([Q_{Po}, Q_{Tr}]), K = Z_{Po}, V = Z_{Po} \tag{4}$$

Using both pose and trajectory queries, past pose embedding Z_{Po} is conditioned on both MPBP sequence at t=0 and the multi-modal trajectory proposals which contain global intent. Subsequently, inverse discrete cosine transform (idct) is applied to convert the future pose proposals from frequency domain to local coordinate domain, $Po_{\mathcal{F}}$. The final multi-modal future pose in global coordinates is acquired as Eq. 5 where $\oplus$ is a joint-wise addition operation.

$$\mathbf{Y} = Tr_{\mathcal{F}} \oplus Po_{\mathcal{F}}, \quad \mathbf{Y} \in \mathbb{R}^{F\times N_A\times T_f\times 3} \tag{5}$$

Table 1. Comparison of statistics between existing human pose forecasting datasets and newly proposed JRDB-GMP dataset.

	Dataset				
	CMU-Mocap (UMPM)	MuPoTs -3D	3DPW	JRDB-GMP	
				1s/2s	2s/5s
Duration (s)	4000	267	1700	1863	
Location #	-	20	-	27	
Sample #	13000	192	432	1153	4593
avg. agent	3	3	2	6.8	6.8
med. agent #	3	3	2	5	5
max agent #	3	3	2	24	22
avg. vel. (m/s)	0.3	0.26	0.57	0.46	0.38
avg. disp.(m)	0.63	0.55	1.13	0.64	0.79
max. disp.(m)	4.62	2.45	10.71	8.44	11.0

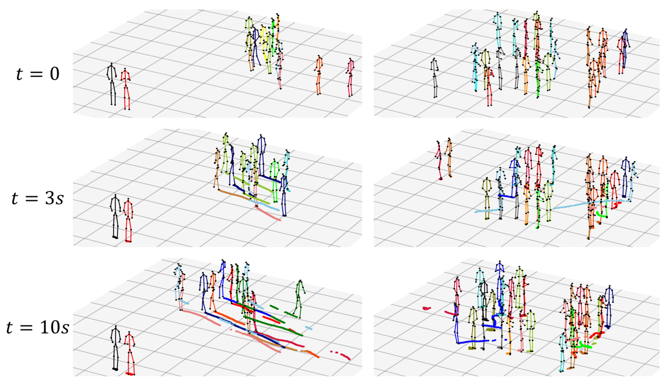

Figure 3. Example scenes from the JRDB-GMP dataset, illustrating its long-term, multi-agent nature.

3.4. Training objective

Both objectives of global trajectory and local pose forecasting are trained jointly. For both global trajectory and local pose prediction, $L2$ loss is propagated to the mode with minimal $L2$ distance with the ground truth.

$$
\begin{gathered}
L_{Tr} = \sum_{n=1}^{N_A} \sum_{t=1}^{T_f} \| \tilde{y}^t_{Tr,n} - \hat{y}^t_{Tr,n} \| \\
L_{Po} = \sum_{n=1}^{N_A} \sum_{t=1}^{T_f} \sum_{j=1}^{J-1} \| (\tilde{y}^{t,j}_{Po,n} - \hat{y}^{t,j}_{Po,n}) \| \\
L = L_{Tr} + L_{Po}
\end{gathered}
\tag{6}
$$

3.5. JRDB-GMP dataset

Due to the absence of existing long-term (3s+) multi-agent (6+) dataset, we compose a unique 3D human pose forecasting dataset in a real-world environment from JRDB [66]. The original JRDB dataset is constructed by a moving robot that records human activity around a school campus using 5 omnidirectional cameras and LiDAR. Image sequences along with 2D pose annotation and 3D bounding box annotations are provided in the original dataset. However, since 3D human pose annotations are unavailable, we separately parse accurate 3D human pose from provided inputs and annotations. First, a SOTA monocular 3D pose extraction method [59] is used to extract raw 3D joint positions from image sequences. Then, 2D pose and 3D bounding box annotations are used to refine the raw joint positions and minimize noise. We use 2D pose annotations to initially filter out the 3D poses with noise. With camera parameters and refined 3D pose, we project it on 2D image plane, then L2 distance between projected 2D pose and GT 2D pose annotation is calculated. If the mean L2 distance per each agent at a time stamp is over a threshold, that instance is filtered out. 2D pose annotations are also projected in 3D space to refine the remaining 3D poses, ensuring the accuracy of the 3D pose information of our JRDB-GMP dataset. Further details are elaborated in the supplementary materials.

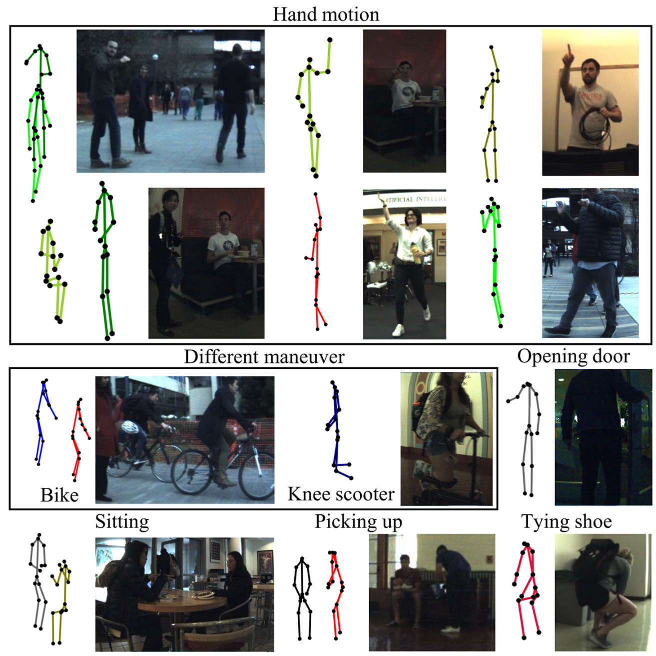

Figure 4. Various motions from the JRDB-MultiGlobPose dataset, providing rich motion queues for inter-agent interaction inference.

Figure 3 visualizes some scenes of the constructed dataset. Accurate extraction of 3D poses has been made even with considerable occlusion via the use of 2D poses. The dataset includes agents with both long and short traverse distances and rich inter-agent interactions in both trajectory and local pose aspects. Figure 4 illustrates diverse local poses included in the dataset, which serve as motion cues of inter-agent interaction. Both figures confirm our method's accuracy in extracting 3D multi-human pose, even in crowded environments. Table 1 scrutinizes the statistics compared to previously used datasets. Compared to earlier datasets, the average number of agents is more than twice as high. In addition, comparing JRDB-GMP 1s/2s to CMU-Mocap and MuPoTs datasets, JRDB contains more diverse and longer motion as shown by a similar magnitude of average displacement but longer maximum displacement.

Table 2. Quantitative comparison of our method to previous methods on CMU-mocap (UMPM), 3DPW, and JRDB-GlobMultiPose datasets with number of prediction modes (F) as 6. Lower is better for all metrics. The best results are marked in **bold**.

Dataset		CMU-mocap (UMPM)		3DPW		JRDB-GlobMultiPose			
In/out length (s)		1/2		0.8/1.6		1/2		2/5	
Evaluation time (s)		1	2	0.8	1.6	1	2	2.5	5
JPE	MRT [69]	164.7	280.1	159.1	251.2	259.3	349.3	438.4	474.0
	JRT [74]	168.5	316.9	181.9	287.3	237.9	373.1	351.9	538.8
	TBIFormer [47]	170.0	290.9	153.9	265.8	257.1	339.3	443.2	481.3
	Ours	**152.4**	**262.7**	**142.6**	**236.2**	**224.0**	**301.4**	**341.6**	**390.4**
APE	MRT [69]	127.0	164.4	117.9	153.2	72.3	87.3	88.5	101.9
	JRT [74]	121.2	181.6	133.4	178.0	112.6	154.3	96.7	120.2
	TBIFormer [47]	125.1	160.8	115.4	152.7	**70.6**	**83.3**	88.2	102.9
	Ours	**114.4**	**151.7**	**114.6**	**150.0**	70.8	**83.3**	**82.2**	**94.7**
FDE	MRT [69]	99.6	204.7	102.7	185.3	235.2	325.2	418.2	454.8
	JRT [74]	117.7	250.8	133.7	235.4	211.4	337.4	318.5	497.2
	TBIFormer [47]	112.1	228.5	106.7	215.9	232.4	314.6	423.9	458.8
	Ours	**88.7**	**188.9**	**74.1**	**158.2**	**194.7**	**271.5**	**313.9**	**361.0**

4. Experiment

4.1. Dataset

We test our model on three datasets: CMU-Mocap (UMPM) [13, 64], 3DPW [67], and our JRDB-GMP. Although our model is designed to forecast human poses in a long-term multi-agent environment, we also report experimental results on previous benchmark datasets with simpler scenes. Mocap-UMPM is a mixed dataset of Mocap and UMPM containing synthesized human interaction between three agents [47]. 3DPW is a dataset with 2 agents traversing a real-world environment. We report the test results on each after separate training on respective datasets.

4.2. Metrics

We use the following widely-used metrics. For a detailed definition, please refer to the supplementary material.

APE: Aligned mean per joint Position Error is used as a metric to evaluate the forecasted local motion. $L2$ distance of each joint in the hip joint coordinate is averaged over all joints for a given timestep.

FDE: Final Distance Error evaluates the forecasted global trajectory by calculating the $L2$ distance of a given timestep.

JPE: Joint Precision Error evaluates both global and local predictions by mean $L2$ distance of all joints for a timestep.

4.3. Implementation details

We train our model on a single A6000 GPU. 2 layers of pose encoder transformer are stacked, followed by 2 layers of transformer in pose decoder. Embedding dimensions of 96 and 128 are used for trajectory and pose embeddings, respectively. The transformed key, value dimension of 64 is used for all transformer architectures. A learning rate of 0.003 is used with an AdamW optimizer with weight decay. Further details can be found in the supplementary materials.

Table 3. Short-term prediction results on CMU-Mocap (UMPM) dataset, where 1s of poses are forecasted given 2s of poses.

Metric	JPE			APE			FDE		
Time (s)	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0
MRT	64.5	152	217	49.8	110	140	39.4	97.9	153
JRT	**31.5**	104	173	**28.7**	85.9	125	17.7	63.9	120
TBIFormer	37.4	104	**158**	32.8	85.8	119	23.3	63.7	104
Ours	37.8	**102**	**158**	33.8	**84.4**	**116**	**14.9**	**49.1**	**92.6**

4.4. Baselines

We compare our method against the latest SOTA methods for multi-agent pose forecasting [47, 69, 74]. To compare the multi-modal predictions of these three methods, we extend their prediction modes by spanning embedding K times in the same manner as ours. All baselines are trained and evaluated on CMU-mocap (UMPM) and 3DPW datasets. CMU-Mocap (UMPM) dataset predicts 2 seconds from 1 second of poses, and 3DPW predicts 1.6 seconds from 0.8 seconds of poses, both from 6 modes. For 3DPW dataset, we slightly lengthen the forecast horizon to evaluate long term predictions. For JRDB-GMP dataset, both short (1s/2s) and long term (2s/5s) predictions are evaluated for all models. Lastly, We use HiVT [88] as the baseline For global trajectory prediction of our model.

5. Results

5.1. Quantitative results

Table 2 compares the quantitative performances on three datasets. Our method exhibits considerable performance gain against all previous SOTA methods, not only on the proposed long-term multi-agent dataset but also on the existing two datasets. Such generalized competence demonstrates the applicability of our trajectory-conditioned pose forecasting method to various real-world scenarios. In detail, our approach achieves over 10% gain of FDE on all datasets. This improvement on forecasting global locomo-

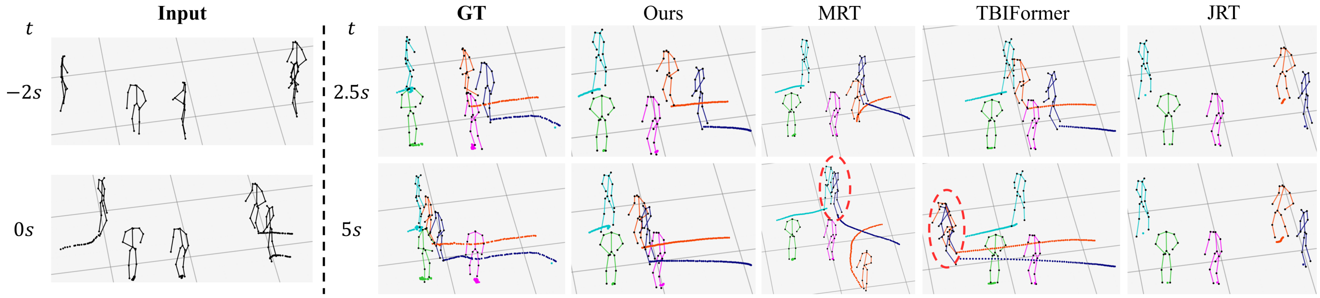

Figure 5. Visualization of a long-term forecasting scene from JRDB-GMP (2/5) dataset. Past poses for input are shown on the leftmost column, GT future poses on the next, and forecasts by ours, MRT, TBIFormer, and JRT, respectively.

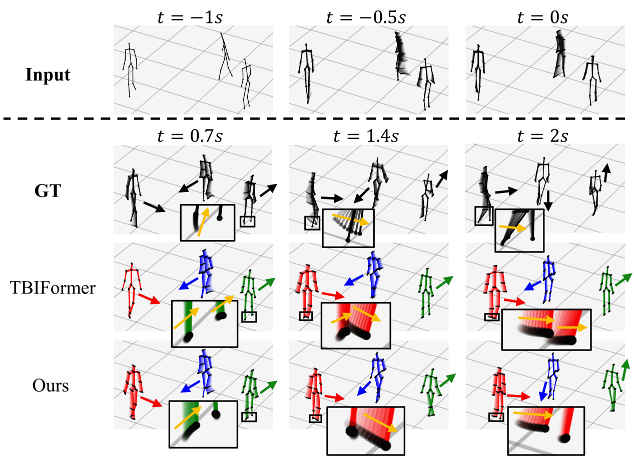

Figure 6. Visualization of a CMU-Mocap (UMPM) scene. Past poses are shown on the upper row, GT future poses on the next, and forecasts by TBIFormer and ours on the latter two rows. To visualize motion, we stack several frames around the target time stamp. Black/red/blue arrows refer to the direction of the global trajectory, and yellow arrows refer to the direction of foot motion.

tion could be accredited to the decoupled forecasting of global trajectory and local pose. Previous methods holistically predict both global and local movements, limiting both performances due to superfluous interactions to consider between all joints. Conversely, our approach can extract accurate global intent by decoupling past motion into global and local representations. Moreover, our effective interaction modeling of global and local pose also helps to predict a more accurate global trajectory under multi-agent environment as shown in the latter ablation study.

For APE metric, our method also surpasses previous SOTA models on all datasets, highlighting the accurate extraction of local pose intent. Such improvement shows that our approach generates plausible local motion due to its proficient sampling from coarse global intents. Our method simplifies the task by learning multi-modality in a coarse-to-fine approach. Its subsequent local motion forecasting is inferred from coarsely modeled multi-modality, a greatly simplified task compared to extracting intent from entangled multi-modality as well as multi-agent interaction.

These improvements on both global and local scales jointly contribute toward lowering the JPE metric, demonstrating proficiency of our method in forecasting overall human motion. Based on such competence, our method which aimed towards improving on forecasting long-term multi-agent environments also exhibits similar or better performances on short timeframes as shown in Tab. 3. In addition, our approach also excels even on sole local motion with minimal global displacement, as elaborated in section 7.2 of supplementary materials.

5.2. Qualitative results

Our method forecasts a more plausible global pose in longer timescales (∼5s) as shown in the interacting scene of five agents in Fig. 5. Looking at the input and GT sequences, the leftmost person avoids the traversing couple from right to left. The two people in front are stationary while talking to each other. MRT and TBIFormer forecasts implausible overlapped poses at the final prediction horizon (t=5s). JRT fails to learn the global locomotion of agents due to the high complexity of its attention mechanism and is stuck in the local minimum of predicting the inactivity of all agents. On the other hand, our model forecasts plausible poses where the closely interacting two agents walk side-by-side.

Figure 6 illustrates exemplary sequences where more natural local motion has been forecasted by our method. Comparing forecasts on a scene of walking agents, our method generates a much more plausible sequence where the stepping foot remains stationary. On the other hand, the previous SOTA method, TBIFormer, struggles to learn the natural walking mechanism of human legs and a parallel translation of both feet is exhibited. Such discrepancy shows that trajectory-conditioning for inferring local motion from global intent generates more proficient details in human motion than SOTA methods. More visualizations could be found in the supplementary materials.

5.3. Ablation studies

Different number of modes. The main quantitative results report prediction results with F as 6 to compare the ability to address the multi-modal nature of human motion during pose forecasting. Table. 4 additionally compares forecast

Table 4. Comparsion of performance with different number of modes in CMU-Mocap (UMPM) dataset.

F	1		6	
Metric @ 2s	APE	JPE	APE	JPE
MRT	163.9	**366.4**	164.4	280.1
JRT	176.7	367.4	181.6	316.9
TBIFormer	160.1	374.3	160.8	290.9
Ours	**154.4**	**366.4**	**151.7**	**262.7**

Table 5. Ablation studies on core components of model structures. Experiments are done with JRDB-GMP dataset to evaluate multi-agent long-term performance.

Exp.	Trajectory encoder		Pose decoder	Metrics		
#	Local pose embedding	Agent interaction	Trajectory -conditioning	JPE @5s	APE @5s	FDE @5s
-				471.4	101.7	457.9
1		✓		400.5	95.1	370.9
2	✓			403.3	94.7	374.2
3			✓	401.2	93.0	372.8
4	✓		✓	395.6	93.8	366.8
5	✓	✓		392.7	95.2	363.4
6	✓	✓	✓	**391.2**	**91.4**	**363.3**

results with F as 1. Our method again achieves noticeable improvement in APE over the baseline on single-modal forecasts. With $F = 1$, although our method barely enjoys improvement in forecasting global motion due to the absence of multi-modality, its superiority in APE shows the validity of our coarse-to-fine forecasting strategy that also effectively captures agent interaction. Our method improves with multi-modal predictions, demonstrating the proficiency of a coarse-to-fine approach in interpreting the stochastic nature of human motion and its intent. Note that our method improves in APE along with an increase in F unlike previous methods, indicating a unique aptitude in addressing the multi-modal nature of not only global locomotion but also local pose intent via trajectory-conditioning.

Importance of each architecture component. Table. 5 reports the influence of core components of our model. For the trajectory encoder, we evaluate the importance of using local pose embedding and modeling agent interaction. Comparing experiments 1, 5 and 3, 4, both show improvements in JPE and FDE metrics with the use of local pose embedding. Our method has taken advantage of detailed local pose cues to infer an agent's global intention. For interaction modeling, its use is beneficial for both global and local forecasts as compared by experiments 4 and 6. These joint improvements demonstrate the importance of considering local and global motion interactions for their respective forecasts. As for the pose decoder, comparisons of experiments 2,4 and 5,6 both show improvements in APE metric. Such consistent improvement verifies the effectiveness of the trajectory-conditioned local motion forecast approach in generating plausible local motion from global intention.

Importance of interaction modeling. Accurate modeling of inter-agent interaction becomes more pivotal to forecast in more complex environments. Indeed, its complexity grows in a long-term multi-agent scene. When holistically considering joint-wise interaction for all timesteps, the computation complexity is acquired as $O(T^2 \cdot N^2 \cdot J^2)$, where T is the number of timesteps, N the number of agents, and J the number of joints. On the contrary, with interaction modeling in global trajectory scale, our method reduces the computation cost by TJ^2 into $O(T \cdot N^2)$. This enables efficient and proficient modeling of intra (pose) and inter (trajectory)-agent interactions as shown by Tab. 6. While body part-wise interaction modeling only improved by 0.52% for TBIFormer, ours improves up to 3.84% with interaction modeling. This demonstrates the proficiency of our efficient interaction modeling-based method in inferring global and local intents from complex interactions. In addition, the gradual improvement of JPE according to a wider interaction range confirms the importance of interaction modeling of more agents, which cannot be learned from the arbitrarily mixed previous datasets.

Table 6. Ablation studies on agent interaction cutoff distance on JRDB-GMP.

	JPE @ 5s	
	TBIFormer	Ours
w/o interaction	483.8	406.0
w/ interaction < 2m	-	403.5
w/ interaction < 4m	-	400.5
w/ interaction all	481.3	**390.4**

6. Conclusion

In this work, we propose a novel interaction-aware trajectory-conditioned approach to handle long-term multi-agent motion forecasting, along with a new dataset suited for such scope. Our proposed model utilizes a coarse-to-fine approach and decouples overall motion prediction into global and local components. Multi-modality of human motion is proficiently modeled via inferring fine local intents from coarse global intents, along with efficient agent-wise interaction modeling. As for the dataset, our JRDB-GMP dataset contains unprecedented long-term (5s+) multi-agent (6+) interactions in a real-world setting. Our method achieves state-of-the-art performance on all previous datasets and JRDB-GMP dataset, offering generalized practical implications in real-world applications.

Acknowledgements This research was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF2022R1A2B5B03002636) and the Challengeable Future Defense Technology Research and Development Program through the Agency For Defense Development (ADD) funded by the Defense Acquisition Program Administration (DAPA) in 2024 (No.912768601).

References

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5(4):6033–6040, 2020. 2

[2] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13390–13400, 2021. 2

[3] Görkay Aydemir, Adil Kaan Akan, and Fatma Güney. Adapt: Efficient multi-agent trajectory prediction with adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8295–8305, 2023. 2

[4] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2317–2327, 2023. 1, 2

[5] Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 791–798. International Joint Conferences on Artificial Intelligence Organization, 2022. Main Track. 2

[6] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020. 2

[7] Wenhao Chai, Zhongyu Jiang, Jenq-Neng Hwang, and Gaoang Wang. Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14655–14665, 2023. 3

[8] Ling-Hao Chen, JiaWei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9544–9555, 2023. 2

[9] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1423–1432. IEEE, 2019. 2

[10] Sehwan Choi, Jungho Kim, Junyong Yun, and Jun Won Choi. R-pred: Two-stage motion prediction via tube-query attention-based trajectory refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8525–8535, 2023. 2

[11] Rohan Choudhury, Kris M. Kitani, and László A. Jeni. Tempo: Efficient multi-view pose estimation, tracking, and forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14750–14760, 2023. 2

[12] Andrea Porfiri Dal Cin, Giacomo Boracchi, and Luca Magri. Multi-body depth and camera pose estimation from multiple views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17804–17814, 2023. 2

[13] CMU-Graphics-Lab. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/, 2003. 2, 6

[14] Runyang Feng, Yixing Gao, Xueqing Ma, Tze Ho Elden Tse, and Hyung Jin Chang. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17131–17141, 2023. 2

[15] Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, and Hyung Jin Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14861–14872, 2023. 3

[16] Xuehao Gao, Shaoyi Du, Yang Wu, and Yang Yang. Decompose more and aggregate better: Two closer looks at frequency representation learning for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6451–6460, 2023. 2

[17] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13041–13051, 2023. 3

[18] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15303–15312, 2021. 2

[19] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15303–15312, 2021. 2

[20] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13053–13064, 2022. 1

[21] Je-Seok Ham, Dae Hoe Kim, NamKyo Jung, and Jinyoung Moon. Cipf: Crossing intention prediction network based on feature fusion modules for improving pedestrian safety. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3665–3674, 2023. 1

[22] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15977–15987, 2023. 2

[23] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 1

[24] Boyuan Jiang, Lei Hu, and Shihong Xia. Probabilistic triangulation for uncalibrated multi-view 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14850–14860, 2023. 2

[25] Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9644–9653, 2023. 2

[26] Zhehan Kan, Shuoshuo Chen, Ce Zhang, Yushun Tang, and Zhihai He. Self-correctable and adaptable inference for generalizable human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5537–5546, 2023. 2

[27] Jeongho Kim, Wooksu Shin, Hancheol Park, and Jongwon Baek. Addressing the occlusion problem in multi-camera people tracking with human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5462–5468, 2023. 1

[28] Mihee Lee, Samuel S Sohn, Seonghyeon Moon, Sejong Yoon, Mubbasir Kapadia, and Vladimir Pavlovic. Muse-vae: multi-scale vae for environment-aware long term trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2221–2230, 2022. 2

[29] Lihuan Li, Maurice Pagnucco, and Yang Song. Graph-based spatial transformer with memory replay for multi-future pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2231–2241, 2022. 2

[30] Huan Liu, Qiang Chen, Zichang Tan, Jiang-Jiang Liu, Jian Wang, Xiangbo Su, Xiaolong Li, Kun Yao, Junyu Han, Errui Ding, Yao Zhao, and Jingdong Wang. Group pose: A simple baseline for end-to-end multi-person pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15029–15038, 2023. 2

[31] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6437–6446, 2022. 1

[32] Takahiro Maeda and Norimichi Ukita. Motionaug: Augmentation with physical correction for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6427–6436, 2022. 1

[33] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15233–15242, 2021. 2

[34] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9489–9497, 2019. 2

[35] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13309–13318, 2021. 1

[36] Wei Mao, Richard I Hartley, Mathieu Salzmann, et al. Contact-aware human motion forecasting. *Advances in Neural Information Processing Systems*, 35:7356–7367, 2022. 2

[37] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Weakly-supervised action transition learning for stochastic human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8151–8160, 2022. 1

[38] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5517–5526, 2023. 2

[39] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130, 2018. 2

[40] Jiquan Ngiam, Vijay Vasudevan, Benjamin Caine, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, David J Weiss, Benjamin Sapp, Zhifeng Chen, and Jonathon Shlens. Scene transformer: A unified architecture for predicting future trajectories of multiple agents. In *International Conference on Learning Representations*, 2022. 2

[41] Daehee Park, Jaewoo Jeong, and Kuk-Jin Yoon. Improving transferability for cross-domain trajectory prediction via neural stochastic differential equation. *arXiv preprint arXiv:2312.15906*, 2023. 2

[42] Daehee Park, Hobin Ryu, Yunseo Yang, Jegyeong Cho, Jiwon Kim, and Kuk-Jin Yoon. Leveraging future relationship reasoning for vehicle trajectory prediction. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[43] Daehee Park, Jaeseok Jeong, Sung-Hoon Yoon, Jaewoo Jeong, and Kuk-Jin Yoon. T4p: Test-time training of trajectory prediction via masked autoencoder and actor-specific token memory, 2024. 2

[44] Sungchan Park, Eunyi You, Inhoe Lee, and Joonseok Lee. Towards robust and smooth 3d multi-person pose estimation from monocular videos in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14772–14782, 2023. 3

[45] Behnam Parsaeifard, Saeed Saadatnejad, Yuejiang Liu, Taylor Mordan, and Alexandre Alahi. Learning decoupled representations for human pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2303, 2021. 2

[46] Qucheng Peng, Ce Zheng, and Chen Chen. Source-free domain adaptive human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4826–4836, 2023. 3

[47] Xiaogang Peng, Siyuan Mao, and Zizhao Wu. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 17121–17130, 2023. 1, 2, 4, 6

[48] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21254–21263, 2023. 3

[49] Muhammad Rameez Ur Rahman, Luca Scofano, Edoardo De Matteis, Alessandro Flaborea, Alessio Sampieri, and Fabio Galasso. Best practices for 2-body pose forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3613–3623, 2023. 2

[50] Dripta S. Raychaudhuri, Calvin-Khang Ta, Arindam Dutta, Rohit Lal, and Amit K. Roy-Chowdhury. Prior-guided source-free domain adaptation for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14996–15006, 2023. 3

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

[52] Saeed Saadatnejad, Ali Rasekh, Mohammadreza Mofayezi, Yasamin Medghalchi, Sara Rajabzadeh, Taylor Mordan, and Alexandre Alahi. A generic diffusion-based approach for 3d human pose prediction in the wild. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8246–8253. IEEE, 2023. 2

[53] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6457–6466, 2022. 1

[54] Tim Salzmann, Hao-Tien Lewis Chiang, Markus Ryll, Dorsa Sadigh, Carolina Parada, and Alex Bewley. Robots that can see: Leveraging human pose for trajectory prediction. *IEEE Robotics and Automation Letters*, 2023. 1, 2

[55] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14761–14771, 2023. 3

[56] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8887–8896, 2023. 3

[57] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14725–14737, 2023. 2

[58] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11209–11218, 2021. 2

[59] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 2, 3, 5

[60] Yuran Sun, Alan William Dougherty, Zhuoying Zhang, Yi King Choi, and Chuan Wu. Mixsynthformer: A transformer encoder-like structure with mixed synthetic self-attention for efficient human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14884–14893, 2023. 2

[61] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4790–4799, 2023. 2

[62] Julian Tanke, Chintan Zaveri, and Juergen Gall. Intention-based long-term human motion anticipation. In *2021 International Conference on 3D Vision (3DV)*, pages 596–605. IEEE, 2021. 2

[63] Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9601–9611, 2023. 2

[64] N.P. van der Aa, X. Luo, G.J. Giezeman, R.T. Tan, and R.C. Veltkamp. Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1264–1269, 2011. 2, 6

[65] Edward Vendrow, Satyajit Kumar, Ehsan Adeli, and Hamid Rezatofighi. Somoformer: Multi-person pose forecasting with transformers. *arXiv preprint arXiv:2208.14023*, 2022. 1

[66] Edward Vendrow, Duy Tho Le, Jianfei Cai, and Hamid Rezatofighi. Jrdb-pose: A large-scale dataset for multi-person pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4811–4820, 2023. 2, 3, 5

[67] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 6

[68] Chenxi Wang, Yunfeng Wang, Zixuan Huang, and Zhiwen Chen. Simple baseline for single human motion forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2260–2265, 2021. 2

[69] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34:6036–6049, 2021. 1, 2, 6

[70] Mingkun Wang, Xinge Zhu, Changqian Yu, Wei Li, Yuexin Ma, Ruochun Jin, Xiaoguang Ren, Dongchun Ren, Mingxu Wang, and Wenjing Yang. Ganet: Goal area network for motion forecasting. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1609–1615. IEEE, 2023. 2

[71] Yucheng Xing and Xin Wang. Hdg-ode: A hierarchical continuous-time model for human pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14700–14712, 2023. 1, 2

[72] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2022. 2

[73] Chenxin Xu, Robby T. Tan, Yuhong Tan, Siheng Chen, Xinchao Wang, and Yanfeng Wang. Auxiliary tasks benefit 3d skeleton-based human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9509–9520, 2023. 2

[74] Qingyao Xu, Weibo Mao, Jingze Gong, Chenxin Xu, Siheng Chen, Weidi Xie, Ya Zhang, and Yanfeng Wang. Joint-relation transformer for multi-person motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9816–9826, 2023. 1, 6

[75] Sirui Xu, Yu-Xiong Wang, and Liangyan Gui. Stochastic multi-person 3d motion forecasting. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[76] Yi Xu, Armin Bazarjani, Hyung-gun Chi, Chiho Choi, and Yun Fu. Uncovering the missing pattern: Unified framework towards trajectory imputation and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9632–9643, 2023. 2

[77] Maosheng Ye, Jiamiao Xu, Xunnong Xu, Tengfei Wang, Tongyi Cao, and Qifeng Chen. Bootstrap motion forecasting with self-consistent constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8504–8514, 2023. 2

[78] Yingxuan You, Hong Liu, Ti Wang, Wenhao Li, Runwei Ding, and Xia Li. Co-evolution of pose and mesh for 3d human body estimation from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14963–14973, 2023. 2

[79] Bruce X.B. Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8818–8829, 2023. 2

[80] Kai Zhai, Qiang Nie, Bo Ouyang, Xiang Li, and Shanlin Yang. Hopfir: Hop-wise graphformer with intragroup joint refinement for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14985–14995, 2023. 2

[81] Yi Zhang, Pengliang Ji, Angtian Wang, Jieru Mei, Adam Kortylewski, and Alan Yuille. 3d-aware neural body fitting for occlusion robust 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9399–9410, 2023. 2

[82] He Zhao and Richard P. Wildes. Where are you heading? dynamic trajectory prediction with expert goal examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7629–7638, 2021. 2

[83] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8877–8886, 2023. 2

[84] Fang Zheng, Le Wang, Sanping Zhou, Wei Tang, Zhenxing Niu, Nanning Zheng, and Gang Hua. Unlimited neighborhood interaction for heterogeneous trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13168–13177, 2021. 2

[85] Jingxiao Zheng, Xinwei Shi, Alexander Gorban, Junhua Mao, Yang Song, Charles R Qi, Ting Liu, Visesh Chari, Andre Cornman, Yin Zhou, et al. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4478–4487, 2022. 1

[86] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency gcn for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6447–6456, 2022. 1

[87] Mu Zhou, Lucas Stoffl, Mackenzie Weygandt Mathis, and Alexander Mathis. Rethinking pose estimation in crowds: Overcoming the detection information bottleneck and ambiguity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14689–14699, 2023. 3

[88] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022. 2, 4, 6

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

[90] Dekai Zhu, Guangyao Zhai, Yan Di, Fabian Manhardt, Hendrik Berkemeyer, Tuan Tran, Nassir Navab, Federico Tombari, and Benjamin Busam. Ipcc-tp: Utilizing incremental pearson correlation coefficient for joint multi-agent trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5507–5516, 2023. 2

[91] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Unsupervised online video object segmentation with motion property understanding. *IEEE Transactions on Image Processing*, 29:237–249, 2019. 1