Multi-Person Pose Forecasting with Individual Interaction Perceptron and Prior Learning

Peng Xiao[1], Yi Xie[1], Xuemiao Xu[1,2,3(✉)], Weihong Chen[1], and Huaidong Zhang[1,2(✉)]

[1] South China University of Technology, Guangzhou, China
[2] Guangdong Engineering Center for Large Model and GenAI Technology
[3] Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information
{xuemx,huaidongz}@scut.edu.cn

Abstract. Human Pose Forecasting is a major problem in human intention comprehension that can be addressed through learning the historical poses via deep methods. However, existing methods often lack the modeling of the person's role in the event in multi-person scenes. This leads to limited performance in complicated scenes with variant interactions happening at the same time. In this paper, we introduce the Interaction-Aware Pose Forecasting Transformer (IAFormer) framework to better learn the interaction features. With the key insight that the event often involves only part of the people in the scene, we designed the Interaction Perceptron Module (IPM) to evaluate the human-to-event interaction level. With the interaction evaluation, the human-independent features are extracted with the attention mechanism for interaction-aware forecasting. In addition, an Interaction Prior Learning Module (IPLM) is presented to learn and accumulate prior knowledge of high-frequency interactions, encouraging semantic pose forecasting rather than simple trajectory pose forecasting. We conduct experiments using datasets such as CMU-Mocap, UMPM, CHI3D, Human3.6M, and synthesized crowd datasets. The results demonstrate that our method significantly outperforms state-of-the-art approaches considering scenarios with varying numbers of people. Code is available at https://github.com/ArcticPole/IAFormer

Keywords: Human Pose Forecasting · Individual Interaction Modeling

1 Introduction

Human Pose Forecasting (HPF) has emerged as a focal point in contemporary research, finding applications across diverse domains such as autonomous driving and human-machine collaboration. The significance of HPF lies in its pivotal role in enhancing machines' understanding of human behavior. By discerning the historical patterns of human actions, machines can proactively deduce future intentions, facilitating improved collaboration with humans.

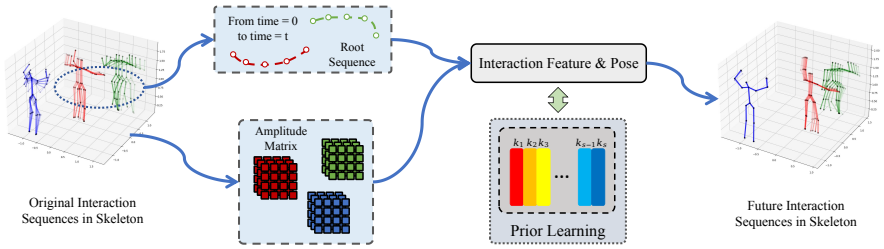

Fig. 1: We extract interaction features from the trajectory and pose amplitude, which reveals the person's role in the event rather than the global information from all the subjects. Moreover, we also find that interaction priors are essential for pose forecasting, so we introduce an interaction knowledge space for prior learning.

Numerous studies have delved into HPF [3, 4, 7, 29]. While these methodologies have shown progress in addressing the correspondence problem. However, human pose development is complex and encompasses various influencing factors. Understanding each person's movements in isolation does not capture the intricacies of interactions within a group, which highlights the need for a more comprehensive approach. Some studies [15, 26, 34] have explored the integration of human interaction within multi-person scenarios, aiming to enhance performance in this challenging task. These approaches treat all individuals equally in a global event, which is too simple to model the scene. In reality, people tend to vary in the degree to which they influence interactions. Unfortunately, these considerations are often absent in current HPF research.

On the other hand, valuable information is embedded in human interaction, exemplified by scenarios where an individual reciprocates a handshake gesture in response to another person's extended hand. Such human gesture responses are ingrained in the human brain as common knowledge in social interactions. Similarly, some tasks benefit significantly from incorporating prior knowledge as a key factor in the model [31, 36]. The accumulation of prior knowledge, specifically social common sense, remains overlooked in current HPF research.

Based on the above insights, we propose a novel architecture IAFormer. As shown in Fig. 1, this work considers the amplitude matrix and root sequence for measuring the person's role in an interaction. To achieve this, we employ the Interaction Perceptron Module (IPM), a key component in the framework. The IPM consists of the Interaction Amplitude Weight constructed based on the Amplitude Matrix of human poses, and the Interaction Trajectory Weight constructed based on the root sequence of individuals. Furthermore, this work also places a strong emphasis on interaction prior learning. The interaction knowledge acquired during the training phase is effectively encapsulated through the Interaction Prior Learning Module (IPLM). The IPLM summarizes and consolidates interaction-related insights to build up Interaction Knowledge Space.

Subsequently, we conducted extensive validations of the IAFormer framework on multiple datasets for Human Pose Prediction experiments, encompass-

ing multi-person and single-person scenarios. Diverse datasets, including CMU-Mocap [5], UMPM [1], CHI3D [13], Human3.6M [16], Mix1 and Mix2 [26], were selected for these experiments. Our contributions can be summarized as follows:

1. We introduce a novel framework, Interaction-Aware Pose Forecasting Transformer (IAFormer). This framework is designed to learn the person-independent interaction features and prior knowledge, providing a general solution to multi-person pose prediction challenges.
2. We present a novel module, the Interaction Perceptron Module (IPM), which utilizes Interaction Amplitude Weight and Trajectory Weight to measure people's role and influence in the interaction.
3. We propose a knowledge-based module, the Interaction Prior Learning Module (IPLM), contributing to learning Human Interaction prior. Notably, our work marks the early attempt at a method incorporating prior knowledge into Human Pose Forecasting.
4. Through extensive experiments on single-person and multi-person datasets, our method outperforms state-of-the-art methods. The demonstrated effectiveness of our approach reaffirms its capability to achieve superior performance across diverse scenarios.

2 Related Work

2.1 Single-Person Human Pose Forecasting

Human Pose Forecasting plays a crucial role in applications like automatic driving, facilitating machines in comprehending human actions. Early endeavors in this domain utilized RNN. Fragkiadaki *et al.* adopted an Encoder-Recurrent-Decoder framework to recognize and predict human body poses in videos and motion capture scenarios [14]. Subsequent advancements by Martinez *et al.* improved the standard RNN model's efficiency for HPF [23]. However, RNN-based models often suffer from error accumulation due to their recurrent nature. Moreover, the HPF task extends beyond being a time-series problem, involving kinematic and anatomical constraints.

Recent efforts [6, 8, 20, 41] have shifted towards convolutional networks such as Graph Convolutional Networks (GCN) [28] and Temporal Convolutional Networks (TCN) to enhance performance by modeling both temporal and spatial information of human pose. Dang *et al.* [8] introduced a multi-scale approach, compressing original human poses and using GCN to extract pose features at different scales. Sofianos *et al.* [30] addressed the lack of understanding of human pose's spatio-temporal dynamics with a Space-Time-Separable Graph Convolutional Network. Ma *et al.* [20] proposed a multi-stage prediction approach for human pose, employing a combination of Spatial Dense Graph Convolutional Networks and Temporal Dense Graph Convolutional Networks to extract spatio-temporal features.

Attention-based methods, inspired by the popularity of Transformers in various tasks [12, 27, 40], have also been employed in recent HPF research. Diller *et*

al. [9] introduced a characteristic pose and used attention-based and volumetric heatmap methods for modeling and generating poses. While these methods exhibit effectiveness in single-person human pose forecasting, their performance diminishes in complex scenarios involving an unlimited number of individuals.

2.2 Multi-Person Human Pose Forecasting

Multi-person scenarios pose a greater challenge than single-person scenarios, given the intricate influence of one's pose on others. Recent advancements in Multi-Person Human Pose Forecasting concentrate on addressing the complexities of human interactions. Adeli *et al*. [2] delve into HPF in wild environments, modeling both human and human-object interactions to predict human trajectory and pose. They employ an attention-based method to extract features from interactions and utilize an RNN-based method for prediction. Guo *et al*. [15] center their attention on extreme collaborative tasks, leveraging cross-attention to model the mutual influence exerted by actors on each other. Meanwhile, Peng *et al*. [26] consider the impact of historical trajectories on HPF, comprehensively modeling both inter-human and human interactions. Although these works exhibit commendable performance and innovation in their respective tasks, they primarily focus on interaction as a key factor while overlooking the individual's role in the interaction. Moreover, these approaches are tailored to scenarios with interactions and often lack experimentation in single-person scenarios.

In this study, we introduce the IAFormer framework, incorporating the Interaction Perceptron Module. Our framework is designed to address both multi-person and single-person scenarios within unlimited-person settings. By considering not only the occurrence of interactions but also the degree of individual involvement, our proposed framework provides a more comprehensive approach to Human Pose Forecasting.

3 Method

Preliminary. We denote the historical pose sequence as $P^i_{1:t} = \{P^i_1, P^i_2, ..., P^i_t\}$ and the sequence of the future poses as $P^i_{t+1:T}$, where i represents the number of humans, t represents the number of the last historical frame and T represents the number of last future frame. To enable the model to learn how human poses change in the future, we follow the [8,20,22] to pad the last historical pose $T-t$ times and append to $P^i_{1:t}$, resulting in the input pose$P^i_{input} = \{P^i_1, ..., P^i_t, ..., P^i_t\}$. Our goal is to predict future pose $P^i_{t+1:T}$ from the padded sequence.

Framework Overview. In IAFormer (as shown in Fig. 2), there is a main transformer branch for specific human pose forecasting and an interaction-aware branch for learning interaction information and prior. The skeleton will be converted to feature space from spatial space through Discrete Cosine Transform (DCT) [22] and Multi-Pose Encoder. Inverse Discrete Cosine Transform (iDCT) and Multi-Pose Decoder will convert the skeleton feature back to spatial space. To extract latent information in feature space, IAFormer employs the Interaction

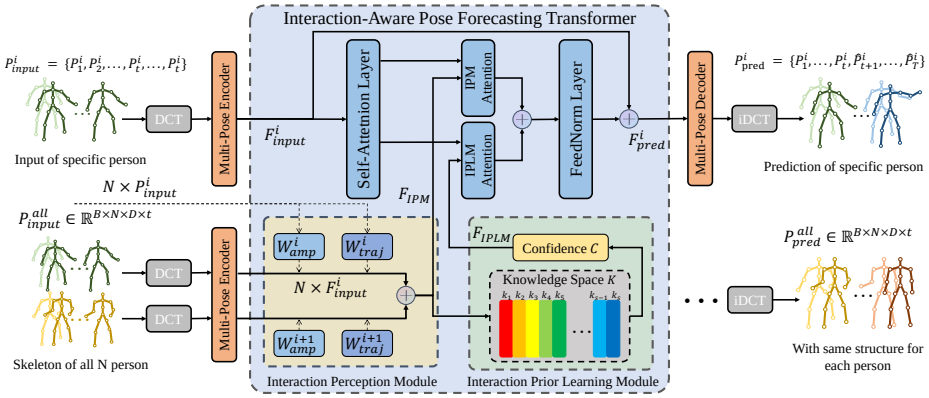

Fig. 2: Overview of Interaction-aware Pose Forecasting Transformer (IAFormer). In IAFormer, given a sequence of poses for several people, the future poses of each person are predicted by combining single-person information and multi-person interaction information separately.

Perceptron Module (IPM) to understand the interaction influence. Additionally, the Interaction Prior Learning Module (IPLM) is utilized to construct interaction knowledge space and assess the reliability of the extracted interaction feature. Further details on these components will be presented in subsequent sections.

Multi-Pose Encoder and Decoder. GCN [18] are particularly well-suited for graphically structured data, excelling at exploring potential connections among nodes [35]. Notably, Mao *et al*. [22] demonstrated strong performance in human pose forecasting by employing multiple GCN Blocks. Inspired by their success, we construct the Multi-Pose Encoder and Decoder in our method, with the GCN Block based on [20] serving as the main body. Such architecture combines DCT and iDCT and is designed for mapping pose information between 3D spatial space and the feature space. For example,

$$F^i_{input} = \text{Encoder}(P^i_{input}), \tag{1}$$

where Encoder$(\cdot)$ represents a function including DCT and Multi-Pose Encoder and F^i_{input} represents the pose feature in the feature space of the No.i human.

3.1 Interaction Perceptron Module (IPM)

This paper predicts human pose by considering the people's role in interaction. In the interaction, each participant has a different influence. For example, the person presenting in a meeting has a higher influence on the entire meeting interaction. This leads to a different weight for each participant in the model.

Therefore, we should quantify the influence of each person in the interaction instead of treating all participants identically. We designed the Interaction Perceptron Module (IPM), which analyzes each person's historical actions to obtain

Amplitude Weight and Trajectory Weight and derives the degree of participation of each person in the current interaction.

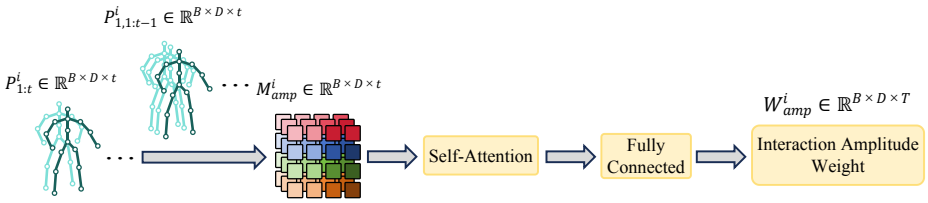

Fig. 3: Interaction Amplitude Weight is constructed by subtracting the motion from the previous frame and mapping function. B means the batchsize, D represents the dimension of the pose also equal to joints $J \times 3$.

Interaction Amplitude Weight. In this paper, we believe that individuals who show more movement within an interaction are more likely to be actively participating and exerting more influence—a phenomenon analogous to people being drawn to moving objects.

Building on this insight, we introduce the concept of Interaction Amplitude Matrix which is denoted as M^i_{amp}, and leverage it as raw information for extracting details about each person's interaction amplitude. We chose to use the coordinate transformation of the previous and next frames of historical pose $P^i_{1:t}$ to obtain the M^i_{amp}. To obtain the subtraction of the last frame from a specific historical frame, we duplicate the first historical frame, creating $P^i_{1,1:t-1} = [P^i_1; P^i_{1:t-1}\}$, which maintains the same size as $P^i_{1:t}$. Then $M^i_{amp} = P^i_{1:t} - P^i_{1,1:t-1}$.

Interaction Amplitude Weight corresponding to the M^i_{amp} is denoted as W^i_{amp} (as shown in Fig. 3). The M^i_{amp} in 3D space cannot fully and directly correspond to the information in the feature space. We design mapping modules for extracting deeper features in $M^i_{amp} \in R^{B\times D\times t}$ and mapping Amplitude Matrix in 3D space to Amplitude Weights $W^i_{amp} \in R^{B\times D\times T}$ in feature space. The mapping modules include 5 Self-Attention Layers and 2 Fully-Connected Layers, which Self-Attention Layers follow the classic design [32]. The W^i_{amp} is calculated as:

$$W^i_{amp} = \mathrm{FC}(\mathrm{SA}(M^i_{amp})), \tag{2}$$

where $\mathrm{SA}(\cdot)$ and $\mathrm{FC}(\cdot)$ corresponds to the Self-Attention and Fully-Connected modules in Fig. 3.

Interaction Trajectory Weight. Grounded in real-world observations, we posit that proximity to the center of a human interaction correlates with active participation and strong influence in this interaction. Accordingly, our proposed Interaction Trajectory Weight W^i_{traj} is designed to reflect a person's degree of dominance in a particular human interaction, which tends to reflect the dominance and influence of an individual within the interaction.

In obtaining W^i_{traj}, we recognize the potential weakness introduced by assessing dominance based solely on the last frame position. Instead, we better

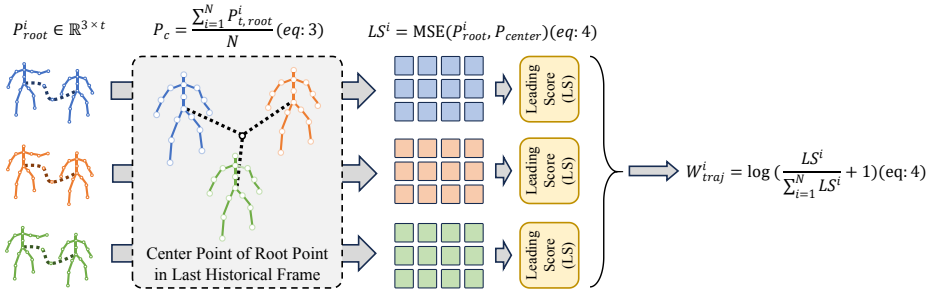

Fig. 4: Interaction Trajectory Weight was calculated by subtracting the center of the interaction from the set of historical root coordinates for each individual.

reflect the person's position changes in the historical sequence by considering all frames. We propose a computational approach for W^i_{traj} illustrated in Fig. 4. This approach focuses on the root point of each pose skeleton which is sufficient to represent global position information in historical poses, denoted as $P^i_{1:t,root} = \{P^i_{1,root}, ..., P^i_{t,root}\}$ and $P^i_{m,root}$ signifies the root coordinate of the m^{th} frame. To compute the center coordinates of the interaction $P_c \in R^{1\times 1}$, we utilize the root coordinates of the last frame for all individuals and calculate as:

$$P_c = \frac{\sum_{i=1}^{N} P^i_{t,root}}{N}. \tag{3}$$

To calculate the $W^i_{traj} \in R^{1\times 1}$, the P_c will be duplicated to the same size as $P^i_{root} \in R^{1\times t}$, which $P_{center} = \{P_c, ..., P_c\} \in R^{1\times t}$. W^i_{traj}'s calculating method is shown as follows:

$$LS^i = \mathrm{MSE}(P^i_{root}, P_{center}), \quad W^i_{traj} = \log(\frac{LS^i}{\sum_{i=1}^{N} LS^i} + 1), \tag{4}$$

where MSE$(\cdot)$ represents the Mean Squared Error function. LS^i is an intermediate variable representing how far an individual's historical trajectory is to the interaction center. When a person is far from the interaction center in the historical period, the corresponding W^i_{traj} will close to zero, and the further from the interaction center the closer to zero.

Individual Information Fusing. This paper focuses on predicting human pose within an interaction, with the view that different individuals within a human interaction contribute to varying degrees. After obtaining W^i_{amp} and W^i_{traj}, IPM applies individual information fusing, adding individual information based on the calculated weight:

$$F_{ipm} = \sum_{i=1}^{N} W^i_{amp} W^i_{traj} F^i_{input}, \tag{5}$$

where F_{ipm} represents the fusion of individual information and F^i_{input} stands for individual information from Multi-Pose Encoder.

3.2 Interaction Prior Learning Module (IPLM)

This paper considers that human interactions encapsulate societal conventions and knowledge. Such knowledge is essential to a machine's comprehension of human social life and is treated as prior in the machine. Individuals develop an understanding of norms and knowledge in social life. To systematically organize and apply interaction knowledge within the model, we design the Interaction Prior Learning Module (IPLM). This mechanism categorizes the extracted interaction information and transforms it into prior. IPLM also provides a corresponding confidence level specific to each kind of interaction.

We built the Interaction Knowledge Space $K = \{k_1, k_2, ..., k_s\}$ based on the Embedding Space mechanism [31]. This mechanism achieves training strong prior on discrete random variables by modeling features in the latent space. Where s represents the space size of K. Our prior learning is achieved by vectorizing the extracted interaction information and categorizing and storing it.

Initialization and Training Phase. As described in Algorithm 1, We first initialize Interaction Knowledge Space K and Learning rate l_r. In the training phase, the input of IPLM is the original interaction feature F_{ipm}. For each input F_{ipm}, IPLM finds the closest vector k_1 in Interaction Knowledge Space K and obtains the loss of K $L_k = \text{MSE}(k_1, F_{ipm})$. Then update the k_1 using F_{ipm}.

Algorithm 1 Interaction Prior Learning Module

INIT Interaction Knowledge Space K, Learning rate l_r
TRAIN(*interaction features* $\mathbf{F_{ipm}}$)
  To $\mathbf{F_{ipm}}$:
    Find the nearest embedding vector k_1 in K
    Update k_1 using sample $\mathbf{F_{ipm}}$ and learning rate l_r:
    $L_k = f_{mse}(k_1, \mathbf{F_{ipm}})$; $k_{1,new} = k_1 + \mathbf{F_{ipm}} * l_r$
  return L_k
PREDICT
  To $\mathbf{F_{ipm}}$:
    Find the first and second nearest embedding vector k_1, k_2 in E
    Save the distance d_1, d_2 between $\mathbf{F_{ipm}}$ and k_1,k_2
    Confidence $\mathbf{C} = 1 - d_1/d_2$
  return $\mathbf{F_{iplm}} = \mathbf{C} * k_1$

Prediction Phase. In the prediction or inference phase, IPLM computes the distance between F_{ipm} and each vector in K and identifies the first and second nearest vectors and corresponding distances. IPLM obtains confidence concerning the F_{ipm} according to $C = 1 - d_1/d_2$, in which d_1 and d_2 respectively indicate the shortest and second shortest distances. The obtained confidence C represents the model's familiarity with the extracted interaction information. IPLM also utilizes the knowledge stored in K to provide a trusted vector F_{iplm} as a feature reference.

3.3 Interaction-Aware Pose Forecasting Transformer

The previous part collects interaction features through IPM and IPLM. We design Interaction-Aware Pose Forecasting Transformer (IAFormer) draws inspiration from the original Transformer architecture [32] for considering both interaction features and individual features. The classic attention formula is shown:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V. \tag{6}$$

A complete IAFormer may include many blocks depending on the difficulty of the task. For a clearer description, we consider one block when introducing IAFormer in the method part, please refer to the experiment section for the architecture details. IAFormer Block includes three procession parts. The first part extracts individual features:

$$I^i_{in} = \text{Attention}(W^Q_1 F^i_{input}, W^K_1 F^i_{input}, W^V_1 F^i_{input}), \tag{7}$$

where W^y_x represents the y weight of x^{th} attention part and I^i_{in} represents the intermediate value output. The second part can fuse individual and interaction features in a two-branch structure. We design to consider both features from IPM and IPLM:

$$\begin{aligned} I^i_{ipm} &= \text{Attention}(W^Q_2 I^i_{in}, W^K_2 F_{ipm}, W^V_2 F_{ipm}), \\ I^i_{iplm} &= \text{Attention}(W^Q_3 I^i_{in}, W^K_3 F_{iplm}, W^V_3 F_{iplm}), \\ I^i_{out} &= 0.5 * I^i_{ipm} + 0.5 * I^i_{iplm}. \end{aligned} \tag{8}$$

Both the IPM Attention and IPLM Attention are shown in Fig. 2. The third procession part is based on Fully Connected Layers and Normalization, denoted as FeedNorm($\cdot$). We also build up a feature residual between the Multi-Pose Encoder/Decoder. Finally the future pose is predicted with the learned features:

$$P^i_{pred} = \text{Decoder}(F^i_{input} + \text{FeedNorm}(I^i_{out})), \tag{9}$$

where Decoder($\cdot$) represents a function including iDCT and Multi-Pose Decoder, P^i_{pred} represents the final predicted pose in 3D space of the No.i human.

3.4 Optimization

In this work, we designed the spatial loss of the model based on the JPE metric:

$$L_s = \frac{1}{J * N} \sum_{i=1}^{N} \sum_{j=1}^{J} \|\hat{P}^i_j - P^i_j\|^2, \tag{10}$$

where $\hat{P}^i_j$ represents the predicted coordinate of the j joint of the No.i human, P^i_j represent the ground truth, N represents the number of humans, J represents the number of body joints.

Temporal consistency is widely utilized across various domains to enhance content generation in time series [39] and is effective in mitigating time-series jitter in generated content. We incorporate temporal consistency by concatenating all frames in the ground truth and prediction. These concatenated frames are then input into the same CNN f_{conv} for feature mapping. The mapped features are used to calculate Mean Squared Error (MSE), yielding the temporal loss:

$$L_t = \text{MSE}(f_{\text{conv}}(P_{pred}), f_{\text{conv}}(P_{gt})). \tag{11}$$

In IAFormer, IPLM updates k_1 in Knowledge Space K and generates a loss L_k.

Compared to previous studies just using the spatial loss, the IAFormer utilizes multiple losses for comprehensive updates:

$$L_{final} = \alpha L_s + \beta L_t + \gamma L_k. \tag{12}$$

4 Experiment

To demonstrate the superiority of our IAFormer method across scenarios involving varying numbers of individuals, we conduct evaluations on six widely used public human pose datasets: CMU-Mocap [5], UMPM [1], CHI3D [13], Human3.6M [16], Mix1 and Mix2 [26]. In what follows, we briefly introduce datasets and performance metrics. Then, we compare IAFormer with state-of-the-art methods. Finally, we conduct ablation experiments.

4.1 Datasets

CMU-Mocap (UMPM) is a mixed 3-person dataset. To ensure that the comparison conditions are consistent, we used the same data CMU-Mocap (UMPM) as [26], which merges UMPM into CMU-Mocap for dataset expansion. In this paper, we set the input frame as 50 and the output frame as 25.

CHI3D is a dataset covering multiple sets of two-person interactions and is suitable for validating the model's ability to perceive interactions. In this paper, we set the input frame as 10 and the output frame as 25.

Human3.6M is one of the most widely used human data sets in the world. In this article, we validate the model's effectiveness in single-person Human Pose Forecasting with this dataset. In this paper, we set the input frame as 10 and the output frame as 25.

Mix1 and Mix2 are a crowded dataset involving a larger number of individuals. Peng et al [26] amalgamated MuPoTS-3D [24], 3DPW [33], and test data from CMU-Mocap and UMPM into two datasets, namely Mix1 and Mix2. The Mix1 dataset comprises 6 individuals and primarily features multi-person interactive motion sequences. On the other hand, the Mix2 dataset involves 10 individuals, encompassing some individuals with minimal or no interaction with others. Each mixed dataset comprises 1000 motion sequences, each including 75 frames.

4.2 Metrics

JPE Metric. We use mean per Joint Position Error (JPE) as the main comparison metric in all experiments, which is a widely used metric in Human Pose Forecasting and Estimation. The calculation formula of JPE is as follows:

$$\mathrm{JPE}(P_{pred}, P_{gt}) = \frac{1}{J * N} \sum_{i=1}^{N} \sum_{j=1}^{J} \|\hat{P}_j^i - P_j^i\|^2. \tag{13}$$

APE Metric. We remove global movement on the skeleton of each human pose and use the Aligned mean per joint Position Error (APE) to measure pure pose position error, which is calculated by:

$$\mathrm{APE}(P_{pred}, P_{gt}) = \mathrm{JPE}(P_{pred} - P_{pred,r}, P_{gt} - P_{pred,r}), \tag{14}$$

where $P_{pred,r}$ and $P_{gt,r}$ are predicted and g round-truth of reference human body.

4.3 Implementation Details

We implement our framework in PyTorch [25], and the experiments are performed on GeForce RTX 4090 GPU. We train our model for 80 epochs using the Adam [10, 11, 17] optimizer with a batch size of 96, and a dropout of 0.1. For optimization, we choose to $\alpha = \beta = \gamma = 1$. There are 5 stacked GCN blocks in the Multi-Pose Encoder/Decoder. There are 5 stacked IAFormer blocks and attention layers with 5 heads in the IAFormer.

In the experiment with CMU-Mocap (UMPM), CHI3D, Mix1 and Mix2, the historical sequence is 50 frames (2s), the predicted sequence is 25 frames (1s), the size of the embedding vector and Interaction Knowledge Space in IPLM is 75 and 256. In the experiment with Human3.6M, a historical sequence is 10 frames (0.4s) and the predicted sequence is 25 frames (1s), size of the embedding vector and Interaction Knowledge Space in IPLM are 35 and 256.

4.4 Comparison with State-of-the-art Methods

To validate the prediction performance of IAFormer, we follow the setting of the most multi-person methods to conduct the comparison experiment between IAFormer and multi-person methods on multi-person datasets (i.e., CMU-Mocap (UMPM) [1,5], Mix1, Mix2 [26], CHI3D [13]). Additionally, to demonstrate the generalizability of our approach, we further conduct a comparative analysis between IAFormer and single-person methods, employing a popular single-person dataset (i.e., Human3.6M [16]) for evaluation. This comprehensive evaluation strategy allows us to assess the performance of IAFormer across different person scenario datasets.

Result on CMU-Mocap (UMPM). From Table 1, we can find that IAFormer outperforms the previous Human Pose Forecasting method. For example, IAFormer exceeds JRFormer [38] by 3mm JPE and 5mm APE in averaging time. IAFormer

Table 1: Performance comparison Results (in mm) on different multi-person datasets. We compare our method with the previous SOTA methods for short-term and long-term predictions. * means multi-person motion prediction method.

	METHOD	CMU-Mocap (UMPM) (3 persons)				Mix1 (6 persons)				Mix2 (10 persons)			
		0.2s↓	0.6s↓	1.0s↓	Avg ↓	0.2s↓	0.6s↓	1.0s↓	Avg↓	0.2s↓	0.6s↓	1.0s↓	Avg↓
JPE	MSR-GCN [8]	53	146	231	143	49	132	220	134	60	153	243	152
	HRI [21]	49	130	207	129	51	141	233	142	52	140	224	139
	MRT* [34]	36	115	193	114	37	122	212	124	38	126	214	126
	TBIFormer* [26]	**30**	109	182	107	34	121	209	121	**34**	118	198	117
	JRFormer* [38]	32	104	161	99	**32**	**109**	**184**	**108**	36	125	211	124
	IAFormer (Ours) *	32	**96**	**159**	**96**	36	112	193	114	36	**108**	**181**	**108**
APE	MSR-GCN [8]	46	106	137	96	41	92	120	84	48	110	148	102
	HRI [21]	41	97	130	89	38	92	122	84	41	100	133	91
	MRT* [34]	36	108	159	101	36	109	166	104	38	115	178	110
	TBIFormer* [26]	27	84	118	76	28	81	113	74	30	89	124	81
	JRFormer* [38]	**20**	78	114	71	**21**	73	105	66	**22**	82	120	75
	IAFormer (Ours)*	23	**71**	**103**	**66**	23	**71**	**101**	**65**	24	**76**	**108**	**69**

Table 2: JPE Results (in mm) on CHI3D [13].

	METHOD	0.2s↓	0.4s↓	0.6s↓	0.8s↓	1s↓	Avg↓
JPE	PGBIG [20]	69	130	181	223	258	172
	TBIFormer* [26]	45	95	145	192	233	142
	IAFormer (Ours)*	**39**	**83**	**129**	**176**	**218**	**129**

outperforms TBIFormer [26] by 11mm JPE and 10 mm APE in averaging time. These results indicate that our method achieves state-of-the-art performance in the 3 person scenario.

Result on Mix1 and Mix2. From Table 1, it is evident that IAFormer maintains state-of-the-art performance even as the number of individuals in the scene increases. This is attributed to the superior feature extraction capabilities of IPM and IPLM, enabling a more effective analysis of complex scene interactions. For example, on Mix1, IAFormer exceeds TBIFormer [26] by 7mm JPE and 9mm APE in averaging time. On Mix2, IAFormer outperforms JRFormer [38] by 1mm JPE and 6mm APE in averaging time. These comparison results demonstrate our method's enhanced capacity to discern complex human interactions, highlighting its unique advantages in addressing such intricate interaction scenarios.

Result on CHI3D. Given that CHI3D involves interactions between two individuals close, it provides an ideal scenario for assessing the efficacy of information extraction at its maximum potential. Consequently, the multi-person human pose forecasting method significantly outperforms single-person human pose forecasting on the CHI3D dataset, owing to its enhanced capability to extract interaction-related information at the highest level. Table 2, TBIFormer [26] outperforms PGBIG [20] by 30.1 JPE in average forecasting time. Since IAFormer can learn from the prior knowledge hidden in the interaction, our IAFormer exceeds TBIFormer [26] by 13.3 JPE. These experiment results show the significant advantage of our method in people with close interaction scenarios.

Table 3: JPE Results (in mm) on Human3.6M [16].

	METHOD	80ms↓	160ms↓	320ms↓	400ms↓	560ms↓	1000ms↓
JPE	DMGNN [19]	17.0	33.6	65.9	79.7	103	137.2
	LTD [22]	2.7	26.1	52.3	63.5	81.6	114.3
	MSR [8]	12.1	25.6	51.6	62.9	81.1	114.2
	PGBIG [20]	10.3	22.7	47.4	58.9	76.9	110.3
	AuxFormer [37]	9.5	20.6	43.4	54.1	75.3	**107.0**
	IAFormer (Ours)	**8.4**	**18.1**	**39.8**	**50.8**	**72.6**	130.4

Table 4: Ablation Studies on CMU-Mocap(UMPM) in JPE.

	IPM (IAW)	IPM (ITW)	IPLM	200ms↓	600ms↓	1000ms↓	Avg↓
Backbone				42.1	112.8	179.1	111.3
With IPM	✓	✓		35.2	100.2	162.2	99.2
With IPLM			✓	35.2	99.3	163.2	99.2
IPLM+IPM(IAW)	✓		✓	35.1	99.7	161.2	98.7
IPLM+IPM(ITW)		✓	✓	33.2	98.7	161.6	97.8
IAFormer (Ours)	✓	✓	✓	**32.1**	**96.5**	**159.2**	**95.9**

Result on Human3.6M. To verify the versatility of IAFormer, we follow the time point setting as the previous work to compare IAFormer with recent single-person methods on a most popular single-person dataset (i.e., Human3.6M [16]), as shown in Table 3. From Table 3, we can find that IAFormer outperforms previous single-person methods. For example, IAFormer outperforms the best single-person method (i.e., AuxFormer [37]) by 3.6mm JPE when forecasting time is 320ms. Furthermore, IAFormer exceeds PGBIG [20] by 8.1mm JPE when forecasting time is 400ms. These results that our method performs well in short-term ($\leq 400ms$) forecasting compared with other single-person methods.

4.5 Ablation Studies

To comprehensively verify the critical role played by each novelty component of the model, we undertake exhaustive ablation experiments using the CMU-Mocap (UMPM) dataset, as shown in Table 4. The "Backbone" case refers to the framework consisting solely of the multi-pose encoder/decoder, devoid of any information from IPM and IPLM.

Effectiveness of IPM. From Table 4, it is evident that the "With IPM" case exceeds the "Backbone" case across all forecasting time. This is because the IPM module is conceived to gauge the influence of each person in human interactions, assessing their amplitude through the intensity of historical actions and dominance through the historical movement trajectories. Specifically, the "With IPM" case outperforms the "Backbone" 16mm JPE in 1000ms. We have also added two parts of IPM to "With IPLM", and these combinations show better performance, which can achieve 1.4mm JPE decrease. These results show the effectiveness of the designed IPM module.

Effectiveness of IPLM. The IPLM module is crafted to comprehend human interaction by constructing the Interaction Knowledge Space. It achieves this by

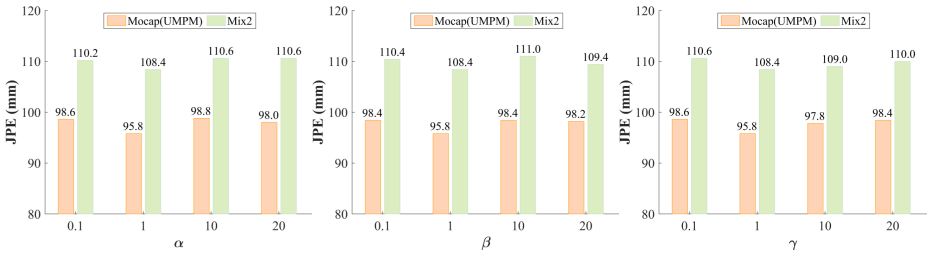

Fig. 5: The influence of three hyper-parameters on average forecasting time.

assessing the degree of understanding of interaction information extracted by the IPM. As depicted in Table 4, the "With IPLM" case outperforms the "Backbone" 12mm JPE in average forecasting time. These results affirm the role of IPLM in helping machines better understand human interaction.

Effectiveness of jointly Employing IPM and IPLM. From Table 4, when IPM and IPLM are used together, the performance of IAFormer is further improved. For example, IAFormer outperforms "With IPM" by 5mm JPE in 1000ms forecasting time. These results show the effectiveness of the proposed IAFormer.

4.6 Hyper-parameter Analysis

The analysis of hyper-parameters α, β, γ in Eq. (12) are conducted in Fig. 5. When a loss corresponds to a small hyper-parameter, it reflects the case of ignoring the loss; while the hyper-parameter is large, it reflects the case of strengthening the effect of the loss. When the three losses are orders of magnitude different from each other, the overall performance of IAFormer decreases. This reflects the rationality of the three losses in the setting.

5 Conclusion

This paper introduces the IAFormer, a multi-person forecasting framework that could learn the potential components prior to interaction. Within IAFormer, the IPM module boosts the efficiency of interaction information extraction by quantifying each person's influence. Additionally, the IPLM module facilitates the accumulation and analysis of commonalities in interaction patterns. Experiments demonstrated that our method outperformed state-of-the-art methods and our method's versatility on multi-person and single-person datasets.

Acknowledgements The work is supported by China National Key R&D Program (Grant No. 2023YFE0202700), Key-Area Research and Development Program of Guangzhou City (No.2023B01J0022), Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (No.2024B1515040010), National Natural Science Foundation of China (No.62302170), Guangdong Basic and Applied Basic Research Foundation (No.2024A1515010187), Guangzhou Basic and Applied Basic Research Foundation (No.2024A04J3750).

References

1. Van der Aa, N., Luo, X., Giezeman, G.J., Tan, R.T., Veltkamp, R.C.: Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 1264–1269 (2011)
2. Adeli, V., Ehsanpour, M., Reid, I., Niebles, J.C., Savarese, S., Adeli, E., Rezatofighi, H.: Tripod: Human trajectory and pose dynamics forecasting in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13390–13400 (2021)
3. Butepage, J., Black, M.J., Kragic, D., Kjellstrom, H.: Deep representation learning for human motion prediction and classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6158–6166 (2017)
4. Chiu, H.k., Adeli, E., Wang, B., Huang, D.A., Niebles, J.C.: Action-agnostic human pose forecasting. In: Proceedings of the IEEE/CVF winter conference on Applications of Computer Vision. pp. 1423–1432 (2019)
5. CMU-Graphics-Lab: Cmu graphics lab motion capture database (2003), http://mocap.cs.cmu.edu/
6. Cui, Q., Sun, H.: Towards accurate 3d human motion prediction from incomplete observations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4801–4810 (2021)
7. Cui, Q., Sun, H., Yang, F.: Learning dynamic relationships for 3d human motion prediction. In: Proceedings of the IEEE/CVF conference on computer vision and Pattern Recognition. pp. 6519–6527 (2020)
8. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: 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 (2021)
9. Diller, C., Funkhouser, T., Dai, A.: Forecasting characteristic 3d poses of human actions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15914–15923 (2022)
10. Ding, Y., Mao, R., Du, G., Zhang, L.: Clothes-eraser: clothing-aware controllable disentanglement for clothes-changing person re-identification. Signal, Image and Video Processing pp. 1–12 (2024)
11. Ding, Y., Wang, A., Zhang, L.: Multidimensional semantic disentanglement network for clothes-changing person re-identification. In: Proceedings of the 2024 International Conference on Multimedia Retrieval. pp. 1025–1033 (2024)
12. Ding, Y., Wu, Y., Wang, A., Gong, T., Zhang, L.: Disentangled body features for clothing change person re-identification. Multimedia Tools and Applications pp. 1–22 (2024)
13. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Three-dimensional reconstruction of human interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7214–7223 (2020)
14. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4346–4354 (2015)
15. Guo, W., Bie, X., Alameda-Pineda, X., Moreno-Noguer, F.: Multi-person extreme motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13053–13064 (2022)

16. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (2013)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
19. Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., Tian, Q.: Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 214–223 (2020)
20. Ma, T., Nie, Y., Long, C., Zhang, Q., Li, G.: 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. pp. 6437–6446 (2022)
21. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: Proceedings of the European Conference on Computer Vision. pp. 474–489 (2020)
22. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9489–9497 (2019)
23. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2891–2900 (2017)
24. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb. In: International Conference on 3D Vision. pp. 120–130 (2018)
25. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems (2019)
26. Peng, X., Mao, S., Wu, Z.: Trajectory-aware body interaction transformer for multi-person pose forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17121–17130 (2023)
27. Shen, F., Xie, Y., Zhu, J., Zhu, X., Zeng, H.: Git: Graph interactive transformer for vehicle re-identification. IEEE Transactions on Image Processing **32**, 1039–1051 (2023)
28. Shen, F., Zhu, J., Zhu, X., Xie, Y., Huang, J.: Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification. IEEE Transactions on Intelligent Transportation Systems **23**(7), 8793–8804 (2021)
29. Shu, X., Zhang, L., Qi, G.J., Liu, W., Tang, J.: Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(6), 3300–3315 (2021)
30. Sofianos, T., Sampieri, A., Franco, L., Galasso, F.: Space-time-separable graph convolutional network for pose forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11209–11218 (2021)
31. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)

32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
33. Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision. pp. 601–617 (2018)
34. Wang, J., Xu, H., Narasimhan, M., Wang, X.: Multi-person 3d motion prediction with multi-range transformers. Advances in Neural Information Processing Systems **34**, 6036–6049 (2021)
35. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.: A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems **32**(1), 4–24 (2020)
36. Xiao, P., Wang, C., Lin, Z., Hao, Y., Chen, G., Xie, L.: Knowledge-based clustering federated learning for fault diagnosis in robotic assembly. Knowledge-Based Systems **294**, 111792 (2024)
37. Xu, C., Tan, R.T., Tan, Y., Chen, S., Wang, X., Wang, Y.: Auxiliary tasks benefit 3d skeleton-based human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9509–9520 (2023)
38. Xu, Q., Mao, W., Gong, J., Xu, C., Chen, S., Xie, W., Zhang, Y., Wang, Y.: Joint-relation transformer for multi-person motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9816–9826 (2023)
39. Zhang, H., Shen, C., Li, Y., Cao, Y., Liu, Y., Yan, Y.: Exploiting temporal consistency for real-time video depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1725–1734 (2019)
40. Zheng, W., Xu, C., Xu, X., Liu, W., He, S.: Ciri: curricular inactivation for residue-aware one-shot video inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13012–13022 (2023)
41. Zhong, C., Hu, L., Zhang, Z., Ye, Y., Xia, S.: Spatio-temporal gating-adjacency gcn for human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6447–6456 (2022)