Multi-Person Extreme Motion Prediction

Wen Guo\textsuperscript{1,2}, Xiaoyu Bie\textsuperscript{1,\*}, Xavier Alameda-Pineda\textsuperscript{1}, Francesc Moreno-Noguer\textsuperscript{2}
\textsuperscript{1}Inria, Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
\textsuperscript{2}Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
\textsuperscript{1}\{wen.guo,xiaoyu.bie,xavier.alameda-pineda\}@inria.fr, \textsuperscript{2}fmoreno@iri.upc.edu

\begin{center}
\includegraphics[width=\textwidth]{figure1.png}
\end{center}

Figure 1. **Collaborative human motion prediction.** 1st row: 3D sample meshes from our ExPI Dataset (just for visualization purposes). 2nd-4th rows: Motion prediction results by MSR \cite{19}, Hisrep \cite{47}, and our method. Dark red/blue indicate prediction results, and light red/blue are the ground truth. By exploiting the interaction information, our approach of collaborative motion prediction achieves significantly better results than methods that independently predict the motion of each person.

\section*{Abstract}

Human motion prediction aims to forecast future poses given a sequence of past 3D skeletons. While this problem has recently received increasing attention, it has mostly been tackled for single humans in isolation. In this paper, we explore this problem when dealing with humans performing collaborative tasks, we seek to predict the future motion of two interacted persons given two sequences of their past skeletons. We propose a novel cross interaction attention mechanism that exploits historical information of both persons, and learns to predict cross dependencies between the two pose sequences. Since no dataset to train such interactive situations is available, we collected ExPI (Extreme Pose Interaction) dataset, a new lab-based person interaction dataset of professional dancers performing Lindy-hop dancing actions, which contains 115 sequences with 30K frames annotated with 3D body poses and shapes. We thoroughly evaluate our cross interaction network on ExPI and show that both in short- and long-term predictions, it consistently outperforms state-of-the-art methods for single-person motion prediction. Our code and dataset are available at: \url{https://team.inria.fr/robotlearn-multi-person-extreme-motion-prediction/}

\section{Introduction}

The goal of human motion prediction is to predict future motions from previous observations. With the successful development of deep human pose estimation from single image \cite{9, 18, 27, 37, 51, 52, 55, 56, 58, 59, 67}, motion prediction begins to draw an increasing attention \cite{3, 8, 16, 22, 23, 26, 29, 33, 38, 43, 47, 49, 50, 60}. Most existing works formulate motion prediction as a sequence-to-sequence task, where past observations of 3D skeleton data are used to forecast future skeleton movements. A common denominator of all these approaches is that they treat each pose
sequence as an independent and isolated entity: the motion predicted for one person relies solely on her/his past motion. However, in real world scenarios people interact with each other, and the motion of one person is typically dependent on or correlated with the motion of other people. Thus, we could potentially improve the performance of motion prediction by exploiting such human interaction.

Based on this intuition, in this paper we present a novel task: collaborative motion prediction, which aims to jointly predict the motion of two persons strongly involved in an interaction. To the best of our knowledge, previous publicly available datasets for 3D human motion prediction like 3DPW [67] and CMU-Mocap [25] that involve multiple persons only include weak human interactions, e.g., talking, shaking hands etc. Here we move a step further and analyse situations where the motion of one person is highly correlated to the other person, which is often seen in team sports or collaborative assembly tasks in factories.

With the goal to foster research on this new task, we collected the ExPI (Extreme Pose Interaction) dataset, a large dataset of professional dancers performing Lindy Hop aerial steps.¹ To perform these actions, the two dancers perform different movements that require a high level of synchronisation. These actions are composed of extreme poses and require strict and close cooperation between the two persons, which is highly suitable for the study of human interactions. Some examples of this highly interacted dataset are shown in Figure 2. Our dataset contains 115 sequences of 2 professional couples performing 16 different actions. It is recorded in a multiview motion capture studio, and the 3D poses and 3D shapes of the two persons are annotated for all the 30K frames. We have carefully created train/test splits, and proposed two different extensions of the pose evaluation metrics for collaborative motion prediction task.

To model such strong human-to-human interactions, we introduce a novel Cross-Interaction Attention (XIA) module, which is based upon a standard multi-head attention [64] and exploits historical motion data of the two persons simultaneously. For a pair of persons engaging in the same activity, XIA module extracts the spatial-temporal information from both persons and uses them to guide the prediction of each other.

We exhaustively evaluate our approach and compare it with state-of-the-art methods designed for single human motion prediction. Note that in our dataset of dancing actions, movements are performed at high speed. The long term predictions are very challenging in this case. Nevertheless, the results demonstrate that our approach consistently outperforms these methods by a large margin, with 10 ∼ 40% accuracy improvement for short (≤ 500 ms) and 5 ∼ 30% accuracy improvement for long term prediction (500 ms ∼ 1000 ms).

Our key contributions can be summarized as follows:
- We introduce the task of collaborative motion prediction, to focus on the estimation of future poses of people in highly interactive setups.
- We collect and make publicly available ExPI, a large dataset of highly interacted extreme dancing poses, annotated with 3D joint locations and body shapes. We also define the benchmark with carefully selected train/test splits and evaluation protocols.
- We propose a method with a novel cross-interaction attention (XIA) module that exploits historical motion of two interacted persons to predict their future movements. Our model can be used as a baseline method for collaborative motion prediction.

2. Related Work

2.1. 3D Human Motion Prediction

Due to the inherent sequential structure of human motion, 3D human motion prediction has been mostly addressed with recurrent models. For instance, Fragiadaki et al. [22] propose an encoder-decoder framework to embed human poses and an LSTM to update the latent space and predict future motion. Jain et al. [33] split human body into sub-parts and forward them via structural RNNs. Martinez et al. [50] introduce a residual connection to model the velocities instead of the poses themselves. Interestingly, they also show that a model trained with diverse action data performs better than those trained with single actions. However, although RNNs achieve great success in motion prediction, they suffer from containing the entire history with a fixed-size hidden state and tend to diverge to a static pose. Some works alleviate this problem by using RNN variants [15, 45], sliding windows [10, 11], convolutional models [29, 30, 39] or adversarial training [26].

Since human body is a non-rigid and structured data, directly encoding the whole body into a compact latent embedding will neglect the spatial connectivity of human joints. To this end, Mao et al. [49] introduces a feed forward graph convolutional network (GCN) [35, 65] with learnable adjacent matrix. This approach was later boosted with self-attention on an entire piece of historical information [47] or a selection of them [41]. Recently, GCN based methods are further developed by leveraging multi-scale supervision [19], space-time-separable graph [63], and contextual information [1, 2]. In terms of GCN design, Cui et al. [17] argue that training the adjacent matrix from scratch ignores the natural connections of human joints, and propose to use a semi-constrained adjacent matrix. Li et al. [42] combine a graph scattering network with a hand-crafted adjacent matrix. Other works also exploit the use of transformers [64] to replace GCN in human motion prediction [3, 12].

¹The Lindy Hop is an African-American couple dance born in the 1930’s in Harlem, New York, see [54].
Considering that human actions are essentially stochastic in the future, some works leverage on generative models (e.g. VAEs and GANs) [5, 6, 13, 48, 57, 70, 71, 73]. Nevertheless, although these models can generate diverse future motions, their prediction accuracy still needs to be further improved when compared to deterministic models.

2.2. Contextual Information in Human Interaction

Humans never live in isolation, but perform continuous interactions with other people and objects. Modeling such interactions and the contextual information has been proven to be effective in the topic of 3D human pose estimation [27, 28, 34, 68, 69, 72]. Contextual information has also been shown to be beneficial in predicting human path trajectories. For this purpose, recent works explore the use of multi-agent context with social pooling mechanisms [4], tree-based role alignment [20], soft attention mechanisms [66] and graph attention networks [31, 36, 40].

Unlike the trajectory forecasting problem that focuses on a single center point, motion prediction aims at predicting the dynamics of the whole human skeleton. Incorporating contextual information in such a situation is still much unexplored. Corona et al. [16] expand the use of contextual information into motion prediction with a semantic-graph model, but only weak human-to-human or human-to-object correlations are modeled. Cao et al. [14] involve scene context information into the motion prediction framework, but without human-to-human interaction. More recently, Adeli et al. [1,2] develop a social context aware motion prediction framework, where interactions between humans and objects are modeled either with a social pooling [1] or with a graph attention network [2]. However, they only study in 2D space [7] or with weak human interactions [67]. Since in this dataset [67], most of the actions involve weak interactions like shaking hands or walking together. In any event, none of these papers explores the situation we contemplate in this paper, in which humans do perform highly interactive actions.

2.3. Datasets

Using deep learning methods to study 3D human pose tasks relies on high-quality datasets. Most previous 3D human datasets are single person [32, 46, 62] or made of pseudo 3D poses [53, 67]. Other datasets which contain lab-based 3D data usually do not have close interactions [25, 44, 53, 61]. Recently, some works start to focus on the importance of context information and propose datasets to model interaction of synthetic persons with scenes [14]. Furthermore, Fieraru et al. [21] created a dataset of human interaction with a contact-detection-aware framework, but this dataset just contains several daily scenarios with mild human interactions and it is not released yet at the time of our submission. Thus, we believe the ExPI dataset we present here, where the actions of people are highly correlated, fills an empty space in the current datasets of human 3D pose/motion.

3. Problem Formulation

As discussed in the introduction, the task of single person human motion prediction is well established. It is defined as learning a mapping \( M : P_{t_1:t-1} \rightarrow P_{t:t_1} \) to estimate the future movements \( P_{t_1:t_1} \) from the previous observation \( P_{t_1:t-1} \), where \( t_1 \) denotes the initial (ending) frame of a sequence, and \( P_t \) denotes the pose at time \( t \).

In this work, we extend the problem formulation to collaborative motion prediction of two interacted persons. While our formulation is general and could work for any kind of interactions, for the sake of consistency throughout the paper, we will denote by \( \ell \) and \( f \) variables corresponding to the leader and the follower respectively (see Section 4 on the dataset description). Therefore, the collaborative motion prediction task is defined as learning a mapping:

\[
M_C : P^\ell_{t_1:t_1-1} \rightarrow P^f_{t_1:t_1}, \quad P^f_{t_1:t_1-1} \rightarrow P^\ell_{t_1:t_1},
\]

(1)

Since the two persons are involved in the same interaction, we believe it is possible to better predict the motion of a person by exploit the pose information of her/his interacted partner. From now on, we will use \( P^c_t = [P^\ell_t, P^f_t] \) to denote the joint pose of the couple (two actors) at time \( t \), and \( P_t \) to denote either of them.

In the following parts of the paper, we will provide an experimental framework for the collaborative motion prediction task, consisting of a dataset and evaluation metrics.
16 different actions are performed in the ExPI dataset.

4.1. Dataset Overview

In this section we firstly describe the recording procedure, the camera setup equipped also with a motion-capture system. More details about the data structure and data post-processing are provided in the supplementary material. To foster research in this direction, we will also introduce our proposed method for this task.

4. The Extreme Pose Interaction Dataset

We present the Extreme Pose Interaction (ExPI) Dataset, a new person interaction dataset of Lindy Hop dancing actions. In Lindy Hop, the two dancers are called leader and follower. We recorded 2 couples of dancers in a multi-camera setup equipped also with a motion-capture system. In this section we firstly describe the recording procedure, then give a comprehensive analysis of our dataset.

4.1. Dataset Overview

Dataset Structure. 16 different actions are performed in ExPI dataset, some by the 2 couples of dancers, some by only one of the couples. Each action was repeated five times to account for variability. More precisely, for each recorded sequence, ExPI provides: (i) Multi-view videos at 25FPS from all the cameras in the recording setup; (ii) Mocap data (3D position of 18 joints for each person) at 25FPS synchronised with the videos.; (iii) Camera calibration information; and (iv) 3D shapes as textured meshes for each frame. Overall, the dataset contains 115 sequences with 30k visual frames for each view point and 60k 3D instances annotated.

Dataset Collection and Post-processing. The data were collected in a multi-camera platform equipped with 68 synchronised and calibrated color cameras and a motion capture system with 20 mocap cameras. When collecting the motion capture data, some points are missed by the system due to occlusions or tracking losses, which is a common phenomena in lab-based interacted Mocap datasets [21]. To overcome this issue and ensure the quality of the data, we spent months to manually label the missing points. More details about the data structure and data post-processing are provided in the supplementary material.

4.2. Data Analysis

Diversity. Similar to Ionescu et al. [32], we analyse the diversity of our dataset by checking how many distinct poses have been obtained. We consider two poses to be distinct, if at least one of the J joints for one pose \( P_{m}^{c} \) is different from the corresponding joint of the other pose \( P_{n}^{c} \), beyond a certain tolerance \( \tau \) (mm):

\[
\max_{j \in [1,J]} \| P_{m,j}^{c} - P_{n,j}^{c} \| > \tau, \tag{2}
\]

where \( m, n \in D \) denote any two poses in the dataset \( D \). Then we define diversity of the dataset as the percentage of distinct poses among all the poses. According to Ionescu et al. [32], the diversity of H3.6M is 24% and 12% when setting the tolerance \( \tau \) to 50 mm and 100 mm, respectively. While the diversities of ExPI for the same threshold values are 52% and 23%, which are much more diverse.

Extremeness. To measure the extremeness of a pose sequence, we first compute the standard deviation (std) over time for each dimension of the \( xyz \)—coordinate for every joint. Then, the extremeness of the joint \( j \) is defined as its maximum per-coordinate standard deviation: \( \varepsilon_{j} = \max \{ \sigma_{x}^{j}, \sigma_{y}^{j}, \sigma_{z}^{j} \} \). Finally, the extremeness of an action is evaluated by computing the percentage of joint extremeness values \( \varepsilon_{n} \) within various intervals \([\varepsilon_{\min}, \varepsilon_{\max}]\). Figure 3 reports the extremeness of ExPI dataset compared to H3.6M in two different ways: (i) a per-action plot reporting extremeness on various color-coded intervals (left); (ii) computing the percentage of joints more extreme than a certain std value (right). From both plots it is clear that the ExPI dataset is significantly more extreme than the H3.6M dataset.

5. Method

We introduce our approach for collaborative motion prediction, aiming to set the first performance baseline to help future developments.

5.1. Pipeline

The idea of our method is to learn two person-specific motion prediction mappings, and to propose a strategy to share information between these two mappings. The possibility to include information from the other person involved...
in the interaction, should push the network to learn a better representation for motion prediction. The overall pipeline is described in Figure 4-left.

For the two single person motion prediction mappings, we draw inspiration from [47], using an attention model for leaning temporal attention w.r.t. the past motions, and a predictor based on Graph Convolutional Network (GCN) [35] to model the spatial attention among joints using an adjacency matrix. The temporal attention model aims to find the most relative sub-sequence in the past by measuring the similarity between the last observed sub-sequence and a set of past sub-sequences. In this attention model, the query $Q_t$ is learnt by MLP from the last observation $P_{t-1-M+1}$ (blue dashed rectangle in Figure 4-left, length $M$). The keys $K_i$ are learnt by MLP from the starting chunk of sub-sequences $P_{t_i:t_i+M}$ (red dashed rectangles in Figure 4-left, length $M$). And the values $V_i$ consist of DCT representations built from the sub-sequences $P_{t_i:t_i+M+T}$ (black dashed rectangles in Figure 4-left, length $M+T$), where $t_i$ with $i \in \{1, \ldots, N\}$ indicates the start frame of each past sub-sequence.

Training such strategy separately for each actor does not account for any interaction between the two dancing partners. To deal with this, we design a cross-interaction attention (XIA) module based on multi-head attention, to introduce guidance from the interacted person. In the next section we introduce this XIA module.

5.2. Cross-Interaction Attention (XIA)

XIA aims to share motion information between the two predictors. In particular, we denote the query and the key-value pairs for one person by $Q$ and $\{K_i, V_i\}_{i=1}^N$ respectively, and use the superscript $f$ and $\ell$ to indicate the two person, follower and leader. We naturally cast the collaborative human motion prediction task into learning how to jointly exploit the information in $(K_i, V_i)$ when querying with $Q$ to predict motion of each person.

Our intuition is that the pose information (key-value pairs) of one person can be used to transform the pose information of the other person for better motion prediction. We implement this intuition with the help of the proposed cross-interaction attention module. Such a module takes as input $w$ and the corresponding vector from the interacted pose $w_{\text{int.}}$, and uses multi-head self attention to get the refined vector $\tilde{w}$ (see Figure 4-right):

$$\tilde{w} = \text{XIA}(w_{\text{int.}}, w) = \text{FC}(\text{MHA}(w_{\text{int.}}, w, w) + w),$$

(3)

where MHA$(q, k, v)$ stands for multi-head attention with query $q$, key $k$ and value $v$, and FC indicates fully connected layers. We use different XIA modules to update keys and values mentioned in Section 5.1: in our implementation, XIA modules for keys have 8 attention heads, and XIA for values have a single attention head. Moreover, we add a skip-connection for the MHA module followed by 2 FC layers. XIA modules for leader/follower do not share weights.

The proposed XIA module is integrated at several stages of the computing flow as shown in Figure 4. More precisely, we refine all keys:

$$\tilde{K}_i^\ell = \text{XIA}(K_i^\ell, K_i^f), \quad \tilde{K}_i^f = \text{XIA}(K_i^f, K_i^\ell),$$

(4)

and analogously for the values.

XIA could be potentially generalised to any number of participants by considering either several XIA modules and fusing their outcome, or performing the fusion at the input of XIA module.
5.3. Pose Normalization

Raw poses of ExPI are represented in world coordinate. Similar with single person motion prediction, we normalize the data by removing the global displacement of the poses based on a selected root joint. While our task aims at predicting not only the the distinct poses but also the relative position of the two persons, so we have to normalize by the leader for better visualization. We could either normalize by leader/follower and we choose to normalize by the leader for better visualisation. Specifically, for each frame, we take the the root joint (middle of the two hips) of the leader as coordinate origin, use the root point and left hip of the leader to define x-axis, and use the neck of leader to determine XOZ plane. We normalize all the joints of both persons to this coordinate, then the pose errors can be computed directly in this coordinate. More precisely, we represent the raw poses in world coordinate as \( P_w \in \{ P^f_{w,t}, P^l_{w,t} \} \), and \( T_{w,t}^l \) is the rigid transformation aligning the two actors to the leader’s coordinate system. The normalized coordinates are thus \( P^f_t = T^l_{w,t} P^f_{w,t} \) and \( P^l_t = T^l_{w,t} P^l_{w,t} \). In the following \( P \) shall always represent the normalized pose unless specified otherwise.

6. Experimental Evaluation

This section describes the experimental protocol on ExPI, and discuss the results of our proposed method.

6.1. Splitting the ExPI Dataset

As described in Sect. 4.1, we record 16 actions in ExPI dataset. Seven of them are common actions (\( A_1 \) to \( A_7 \)) which are performed by both of the 2 couples: we denote them as \( A^l_1 \) performed by couple 1 and \( A^l_2 \) by couple 2. The other actions are couple-specific, which are performed only by one couple: we denote the actions performed by couple 1 (\( A_8 \) to \( A_{13} \)) as \( A^l_8 \), and actions by couple 2 (\( A_{14} \) to \( A_{16} \)) as \( A^2_8 \). With these notations, we propose three data splits.

**Common action split.** Similar to [32], we consider the common actions performed by different couples of actors as train and test data. More precisely, \( A^l_2 \) is the train dataset and \( A^l_1 \) is the test dataset. Thus, train and test data contain the same actions but performed by different people.

**Single action split.** Similar to [22, 33], we train 7 action specific models separately for each common action, by taking one action from couple 2 as train set and the related one from couple 1 as test set.

**Unseen action split.** The train set is the entire set of common actions \( \{ A^l_1, A^l_2 \} \). We regard the extra couple-specific actions \( \{ A^l_8, A^2_8 \} \) as unseen actions and use them as our test set. Thus the train and test data contain both couples of actors, but the test actions are not used in training.

To sum up, common action split is designed for a single model on different actions, single action split is designed for action-wise models, and unseen action split focuses on testing unseen actions to measure methods generalization.
### 6.2. Evaluation Metrics

The most common metric for evaluating 3D joint position in pose estimation and motion prediction tasks is the mean per joint position error \( \text{MPJPE}(P,G) = \frac{1}{J} \sum_{j=1}^{J} \| P_j - G_j \|_2 \), where \( J \) is the number of joints, \( P_j \) and \( G_j \) are the estimated and ground truth position of joint \( j \). Based on MPJPE, we propose two different metrics to evaluate the multi-person motion task.

**Joint mean error (JME):** We Propose Joint Mean per joint position Error to measure poses of different persons in the same coordinate, and denote it as JME for simplicity:

\[
\text{JME}(P,G) = \text{MPJPE}(P,G),
\]

where \( P \) and \( G \) are normalized (see Section 5.3) prediction and ground truth. JME provides an overall idea for the performance of collaborative motion prediction by considering the two interacted persons jointly as a whole, measuring both the error of poses and the error of their relative position.

**Aligned mean error (AME):** We Propose Aligned Mean per joint position Error to measure pure pose error without the position bias. We first erase the errors on the relative position between the two persons by normalizing the poses independently to obtain \( \hat{P}, \hat{G} \). However the precision of \( \hat{P} \) is importantly influenced by the joints that are used to determine the coordinate (hips and back). To mitigate this effect, we compute the best rigid alignment \( T_A \) between the estimated pose and the ground-truth using Procrustes analysis [24]:

\[
\text{AME}(P,G) = \text{MPJPE}(T_A(\hat{P}, \hat{G}), \hat{G}),
\]

where \( \hat{P} \in [\hat{P}_l, \hat{P}_f] \) are independently normalized predictions \( \hat{P}_l = T_{\hat{P}, P_l} P_l \) and \( \hat{P}_f = T_{\hat{P}, P_f} P_f \), and \( T_P \) is the normalisation transformation computed from the pose \( P \) as defined in Section 5.3. The same calculation is done for the ground truth \( G \). This normalization is only used for evaluation purpose.

### 6.3. Implementation Details

Since this is the first time the collaborative motion prediction task is presented in the literature, there are no available methods to compare with. Thus we choose 4 code-released state-of-the-art methods of single person motion prediction [19, 47, 49, 50], and implement their released codes\(^5\) on ExPI dataset. For fair comparison, all these models are trained with 50 frames of input, train/test for the leader and the follower separately.

We train our model for 25 epochs and calculate the average MPJPE loss of 10 predicted frames. As the data is normalized by the leader, the corresponding branch converges faster, so we compensate by exponentially down-weighting the loss of the leader with the number of epochs \( \epsilon \), using the loss function: \( L = L_j + 10^{\epsilon} L_{\hat{J}} \).

When predicting longer horizons, we use the predicted motion as input to predict future motion. Inspired by [47], we take 64 sub-sequences for each sequence to reduce the variance of the test results. Overall, we have 7k and 2.3k sub-sequences for training and testing respectively in the common action split and the single action split, and 12k / 2.9k training/testing samples in the unseen action split.

### 6.4. Results and Discussion

**Common action split.** Table 1 reports the results on the common action split. We observe that our proposed method outperforms other methods systematically almost for all actions, in all metrics and for different testing time. In Figure 5-left we calculate the percentage of improvement of our method compared with the state-of-the-art methods, and find that we significantly surpass these methods up to 10 ~ 40% on short term, and 5 ~ 30% on long term. We further compare our per-joint results with Hisrep [47] and MSR [19] in Figure 5-right, and observe that our proposed method gets better results on almost on all the joints. More importantly, the keypoints of the limbs (joints of arms and legs) are improved largely. This is reasonable as interaction between persons comes mostly through the limbs, while joints on the torso have little influence on it. So our

\[^5\]All the codes we use are under MIT license.

---

Table 2. Results on single action split with the two evaluation metrics (in \( \text{mm} \)). Lower value means better performance. Seven action-wise models are trained independently. Our method performs the best in 5 actions, and close to the best for the other 2 actions.
cross-interaction attention is able to improve the accuracy on the limbs more than on the torso. We could also notice the large improvement on the feet of the follower which is closer to the ground truth than the other methods, and it indicates training leader and follower in parallel using our defined loss without XIA module: ‘XIA kqv’ / ‘k’ / ‘kv’ / ‘v’ use XIA module to update key, value and query of the temporal attention, or just some of them.

**Single action split and unseen action split.** We also reported our proposed method by reporting the results on single action split and unseen action split. For single action split, XIA outperforms the state-of-the-art methods also on action-specific models, as shown in Table 2. Interestingly, we observe that the performance on single action split is worse than the corresponding results on common action split, meaning that training on different actions helps regularising the network for this very challenging collaborative extreme motion prediction task. Regarding unseen action split shown in Table 3, we can see that XIA still outperforms the other methods totally fail (Figure 1-right). More qualitative examples could be found in the supplementary material.

**Ablation study.** Taking Hisrep [47] as example, we first tried 3 different ways of training the single-person motion prediction models on our multi-person dataset: (i) ‘mix’: train a single model using data of the two poses \{P^l, P^f\}; (ii) ‘cat’: concatenate the two poses as a single input vector \{P^l, P^f\}; (iii) ‘sep’: train two person-specific models for \{P^l\} and \{P^f\}. Since ‘sep’ gives best performance, all the state-of-the-art methods reported above in this paper is using this setting. As for our collaborative motion prediction model, we report performances of several different design choices of our model. We found that updating the key and values of the temporal attention using our XIA module provide the best results. We demonstrate the interest of the design of our method as the proposed one is the best in performance and our method significantly improves all the single-person motion prediction results.

**Limitations.** Collecting clean and reusable 3D pose data requires specific equipment and recording extreme poses requires actors with specific skills, thus ExPI is rare and difficult to reproduce/extend. This is clearly a limitation in the era of data-hungry deep learning architectures. Besides, predicting very long future (beyond 2s for example) is still an open problem, specially for the fast movements of ExPI.

**7. Conclusion**

Current motion prediction methods are restricted to single person. We move beyond existing approaches for 3D human motion prediction by considering a scenario with two persons performing highly interactive activities. We collected a new dataset called ExPI of professional actors performing dancing actions. ExPI is annotated with sequences of 3D body poses and shapes, opening the door to not only being applied for interactive motion prediction but also for single-frame pose estimation or multi-view 3D reconstruction. In order to learn the interacted motion dynamics, we introduce a baseline method trained with ExPI that exploits historical information of both people in an attention-like fashion. Results of our method show consistent improvement compared to methods that independently predict the motion of each person.
References


[19] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11467–11476, October 2021. 1, 2.1, 1, 2.6, 3.4, 6.4


[73] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for ex-