Expressive Forecasting of 3D Whole-Body Human Motions

Pengxiang Ding[1,2], **Qiongjie Cui**[3,5*], **Haofan Wang**[5], **Min Zhang**[1,2], **Mengyuan Liu**[4], **Donglin Wang**[1]
[1]MiLAB, Westlake University
[2]Zhejiang University
[3]Nanjing University of Science and Technology
[4]Shenzhen Graduate School, Peking University
[5]Xiaohongshu Inc.
dingpengxiang@westlake.edu.cn, cuiqiongjie@njust.edu.cn

Abstract

Human motion forecasting, with the goal of estimating future human behavior over a period of time, is a fundamental task in many real-world applications. However, existing works typically concentrate on predicting the major joints of the human body without considering the delicate movements of the human hands. In practical applications, hand gesture plays an important role in human communication with the real world, and expresses the primary intention of human beings. In this work, we are the first to formulate a whole-body human pose forecasting task, which jointly predicts the future body and hand activities. Correspondingly, we propose a novel Encoding-Alignment-Interaction (EAI) framework that aims to predict both coarse (body joints) and fine-grained (gestures) activities collaboratively, enabling expressive and cross-facilitated forecasting of 3D whole-body human motions. Specifically, our model involves two key constituents: cross-context alignment (XCA) and cross-context interaction (XCI). Considering the heterogeneous information within the whole-body, XCA aims to align the latent features of various human components, while XCI focuses on effectively capturing the context interaction among the human components. We conduct extensive experiments on a newly-introduced large-scale benchmark and achieve state-of-the-art performance. The code is public for research purposes at https://github.com/Dingpx/EAI.

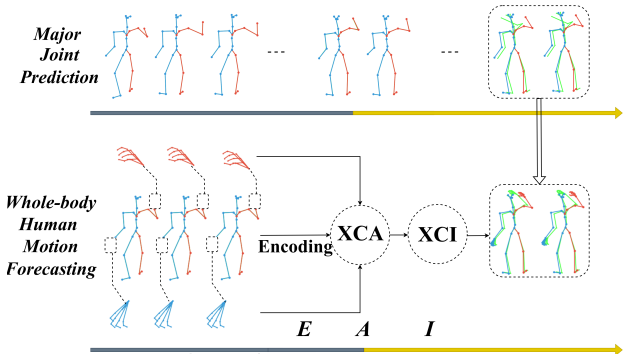

Figure 1: Top: Previous works focus on predicting the human major joints, without considering delicate hand movements that are critical to the HRI application. Bottom: To fill this gap, our work proposes a novel task: whole-body human pose forecasting, to jointly predict future both body and gesture activities. We also highlight that within the proposed EAI, both coarse- (major joints) and fine-grained (gestures) properties are cross-facilitated to achieve a higher-fidelity prediction. Here, red/blue pose is the predicted result, while the underlying green is the ground truth.

Introduction

Predicting the evolution of human behavior/activity in the physical world over time is an essential aspect of machine intelligence (Tarvainen and Valpola 2017; Ruiz, Gall, and Moreno-Noguer 2018; Yuan and Kitani 2020). For instance, to make the seamless human-robot interaction (HRI), a robot is supposed to have some notion of how people will move or act in the near future, conditioned on a series of historically observed poses (Gui et al. 2018; Cui and Sun 2021; Zhang, Black, and Tang 2021; Mao et al. 2019; Cai et al. 2020; Dang et al. 2021; Ding and Yin 2021).

Over the past few years, this attractive topic has received considerable attention, emerging a large number of approaches, with deep learning techniques proving to be sought-after (Cai et al. 2021; Feng et al. 2021; Li et al. 2022; Petrovich, Black, and Varol 2021; Ruiz, Gall, and Moreno-Noguer 2018; Vaswani et al. 2017). Moreover, we note that existing works fall into the coarse-grained scope, *i.e.*, forecasting major joint movements of the human body (Adeli et al. 2021; Cui and Sun 2021; Butepage et al. 2017; Ruiz, Gall, and Moreno-Noguer 2018; Zhong et al. 2022; Ma et al. 2022). However, in terms of realistic applications, it remains a significant limitation: the subtle activity (*i.e.*, gestures) is not considered. The human hand is a vital bridge for interacting with the world, and meanwhile, for the HRI application, it typically includes a detailed command to the robot, embodying human behaviors with the major body (Zhang, Black, and Tang 2021; Diller, Funkhouser, and Dai 2022; Jin et al. 2020; Hidalgo et al. 2019; Taheri et al. 2020). From real applications of human pose forecasting, we, therefore, suggest that considering only major joints, while ignoring the subtle hand gestures, is not sufficient.

To fully investigate this issue, we propose a novel paradigm: *whole-body human motion forecasting*, that is,

*Corresponding author

conjointly predicting future activities of all joints within the body and hands, as shown in Figure 1. In contrast to the conventional task, it presents significant challenges in the following aspects: 1) There are distinct motion patterns in major body and gesture (amplitude of movement, skeletal freedom), and hence it is sub-optimal to treat them equally; 2) A human activity usually involves collaboration/interactivity of different parts within the whole-body; For instance, the clapping-hand embodies the interaction of both hands; and for drinking, it is dominated by the semantic association of the hands and mouth. 3) Due to the heterogeneous scales and characteristics, it is not feasible to directly model such cross-grained interaction as existing multi-person interactive forecasting methods do (Guo et al. 2022).

In this work, we propose a novel Encoding-Alignment-Interaction (EAI) framework to address these challenging issues. Specifically, to avoid negative mutual-interference, we first extract their separate internal spatio-temporal correlations from the body and gesture's heterogeneous motion properties. We observe that, the interaction/collaboration of various elements within the whole-body is critical for performing a specific activity. However, such interaction is incompatible with the existing multi-person interaction (Guo et al. 2022), because person-to-person information is scale-uniform, whereas intra-body context is heterogeneous, *e.g.*, coarse-to-fine-grained (body-to-gesture), or vice versa. Therefore, we propose to exploit the cross-context alignment (XCA) to effectively align and smooth the latent features of different parts, thus eliminating their heterogeneity. Finally, with the aligned features, we further introduce cross-context interaction (XCI), a variant of cross-attention (Hao et al. 2017), that is able to capture the pairwise interactivity between various human parts within the whole-body. We note that, the proposed EAI is a generic framework capable of simutaneously consider the interactivity of different parts within the whole-body as well as the heterogeneous properties, resulting in the higher-quality whole-body prediction.

Our contributions are as follows: (1) To the best of our knowledge, this work is the first to predict the future actions of major joints and human gestures simultaneously. (2) We propose a Encoding-Alignment-Interaction (EAI) approach, equipped with the XCA and XCI, which is capable of extracting the heterogeneous interaction within the whole body. (3) Extensive experiments show that our model achieves superior performance for both short- and long-term prediction compared to the competitors.

Related Work

Human Motion Forecasting. RNNs are the widely-used architectures for time-series data modeling and human pose prediction (Butepage et al. 2017; Bütepage, Kjellström, and Kragic 2018; Honda, Kawakami, and Naemura 2020; Corona et al. 2020). Despite encouraging progress, they typically suffer from error accumulation and tend to converge to a static pose. Feed-forward networks, such as convolutional neural networks (CNNs) (Liu et al. 2020; Ding and Yin 2022) and graph neural networks (GNNs)(Mao et al. 2019; Mao, Liu, and Salzmann 2020; Li et al. 2021, 2022), are proposed as an alternative solution to alleviate the drawbacks of recurrent models. (Mao et al. 2019) presents the learnable adjacent matrix to model spatial dependencies among human body joints. This approach is later extended with self-attention on an entire piece of historical information (Mao, Liu, and Salzmann 2020) or a selection of them (Li et al. 2021). Despite the promising performance, the existing approaches all fall into the scope of predicting the motions of major human joints, without co-analyzing the hand gestures (Pavlakos et al. 2019; Cui and Sun 2021). From the realistic application, we note that subtle hand movements are indispensable for the expressive human behavior and intention. Our work first notices this challenging issue, *i.e.*, how to predict the expressive whole-body human motion (unifying the gesture and major human joints), and aims to solve it.

Contextual Interaction. Contextual Interaction have proven to be effective in the human-to-human interactions (Guo et al. 2022; Wang et al. 2021; Rong, Shiratori, and Joo 2021). Specifically, (Wang et al. 2021) models the context of individual motion and social interactions through a Multi-Range Transformers structure. (Guo et al. 2022) explores the multi-person contextual interactions via a designed cross-interaction module. However, the interaction/-collaboration of various components within the whole-body is incompatible with the above method because human-to-human information is scale-uniform, whereas the intra-body context is heterogeneous. Therefore, in our proposed EAI, we introduce the alignment of heterogeneous features across human components to extract subsequent whole-body internal interactivity more effectively.

Proposed Method

Problem Setup. Previous works typically consider the forecasting of the major human joints. Given T history human poses $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T]$, informally, its objective can be defined as learning a mapping $\mathcal{M} : \mathbf{X} \rightarrow \mathbf{Y}$ to estimate the future poses $\mathbf{Y}$, where $\mathbf{X}$ is the observed major joints, $\mathbf{Y} = [\mathbf{x}_{T+1}, \mathbf{x}_{T+2}, ..., \mathbf{x}_{T+\Delta T}]$ is the corresponding future ones over ΔT frames. This work extends the above standard setup to united whole-body human motion forecasting, including major body, left and right hand, denoted by m, l, r variables for the sake of simplicity. Analogously, we define the novel task as learning a united mapping $\mathcal{M}_{\text{WB}}$:

$$\mathcal{M}_{\text{WB}} : \{\mathbf{X}_l, \mathbf{X}_m, \mathbf{X}_r\} \rightarrow \{\mathbf{Y}_l, \mathbf{Y}_m, \mathbf{Y}_r\} , \qquad (1)$$

where $\mathbf{X}_m \in \mathbb{R}^{D_m \times T}$ ($\mathbf{Y}_m \in \mathbb{R}^{D_m \times \Delta T}$) is the past (future) skeletal sequence of major body. $D_m = 3N_m$ is the number of 3D joint coordinates in a single frame and N_m is the number of major body joints. Also, $\mathbf{X}_l$ ($\mathbf{X}_r$) and $\mathbf{Y}_l$ ($\mathbf{Y}_r$) are the past and future motion of the left (right) hands.

Intra-context Encoding

Due to the distinct motion patterns of major body and gestures, we should consider the different body parts individually. Notably, we extract the intra-context of 3D skeletal sequences comprising of left hand, head, and right hand positions, denoted as $\{\mathbf{X}_l, \mathbf{X}_m, \mathbf{X}_r\}$, in feature space. This is because the spatio-temporal correlations in feature space are more expressive than the correlations in the original motion

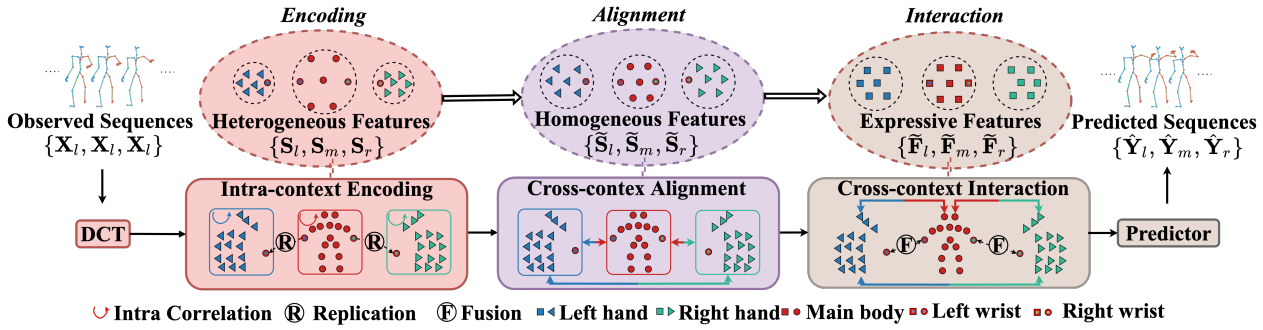

Figure 2: Overall framework of encoding-alignment-interaction (EAI). Given the observed whole-body sequences $\{X_l, X_m, X_r\}$, we first achieve the heterogeneous features $\{S_l, S_m, S_r\}$ via intra-context encoding for each body component independently. Since those intra-context lacks the interaction information of components, the cross-context alignment (XCA) and the cross-context interaction (XCI) are the subsequently proposed to extract cross-context information, where the former aims to alleviate the heterogeneity of components to generate homogeneous features while the latter is designed to explore the cross-context interaction according to the homogeneous features $\{\widetilde{S}_l, \widetilde{S}_m, \widetilde{S}_r\}$ from the XCI. The resulting expressive features $\{\widetilde{F}_l, \widetilde{F}_m, \widetilde{F}_r\}$ are then used to predict future whole-body sequences $\{\hat{Y}_l, \hat{Y}_m, \hat{Y}_r\}$.

space. Next, we illustrate the details of the encoding process by taking the major body as a specific example.

In the temporal domain, Discrete Cosine Transform (DCT) is exploited to capture the temporal smoothness by transforming the observed sequence into trajectory space. Given the past motion $\mathbf{X}_m$, we compute the DCT coefficients of this sequence $\mathbf{X}_m^{''} \in \mathbb{R}^{D_m \times H_c}$ as:

$$\mathbf{X}_m^{''} = \mathbf{X}_m^{'}\mathbf{C}\ , \tag{2}$$

where $\mathbf{X}_m^{'} \in \mathbb{R}^{D_m \times (T+\Delta T)}$ is a variant of $\mathbf{X}_m$ by replicating the last observed pose ΔT times following (Mao et al. 2019); $\mathbf{C} \in \mathbb{R}^{(T+\Delta T)\times H_c}$ is the predefined DCT matrix and each row of $\mathbf{C}$ is the DCT coefficients for a trajectory.

In the spatial domain, we exploit GCNs (Mao et al. 2019; Cui and Sun 2021) to denote the skeleton as a fully-connected graph, depicted as an adjacency matrix $\mathbf{A}_m \in \mathbb{R}^{D_m \times D_m}$. Formally, we define $\mathbf{S}_m^{(n)} \in \mathbb{R}^{D_m \times F^{(n)}}$ as the input feature of n-th layer in GCNs, and $\mathbf{W}_m^{(n)} \in \mathbb{R}^{F^{(n)} \times F^{(n+1)}}$ as the weight matrix. Then the output feature $\mathbf{S}_m^{(n+1)}$ is derived as:

$$\mathbf{S}_m^{(n+1)} = \sigma(\mathbf{A}_m^{(n)}\mathbf{S}_m^{(n)}\mathbf{W}_m^{(n)})\ , \tag{3}$$

where $\mathbf{S}_m^{(1)} = \mathbf{X}_m^{''}$ is the input feature and $F^{(1)} = H_c$ at the first layer; The number of hidden layers $F^{(n)}$ are set to H_d; $\sigma(\cdot)$ is an activation function. The final output features of last layer are $\mathbf{S}_m^{(last)} \in \mathbb{R}^{D_m \times H}$, w.r.t $\mathbf{S}_m$.

Following the above formalism, we also attain the intra-context of the left (right) hands, forming the whole-body intra-context $\{\mathbf{S}_l,\ \mathbf{S}_m,\ \mathbf{S}_r\}$. We note that, although in the standard anatomy the wrist is considered to come from the major body, due to the physical connection with hands, we also include it in the hands feature extraction. Accordingly, the feature dimension of the hands is slightly changed as $\mathbf{S}_l \in \mathbb{R}^{(D_l+3)\times H}$ and $\mathbf{S}_r \in \mathbb{R}^{(D_r+3)\times H}$.

Cross-Context Alignment (XCA)

In contrast to person↔person interactivity (Guo et al. 2022) task, the intra-body context of different body component is scale-inconsistent due to the distinctive motion patterns. The intra-body context is heterogeneous, coarse-to-fine-grained (body-to-gesture), or vice versa, in contrast to the current multi-person interaction where person-to-person information is scale-uniform. In other words, these differences in motion patterns need to be alleviated because they may enormously disturb the overall motion perception.

Intuitively, the heterogeneous context within the whole-body mainly originated from the amplitude of movement, scale, and skeletal freedom of different parts, typically reflecting the discrepancy of feature distribution (Li et al. 2018). To address this issue, we introduce cross-neutralization among the body components combined with discrepancy constraints to effectively align the latent features of different parts. Specifically, we neutralize the distribution discrepancy among different features via a learnable factor that can be automatically adjusted according to the MMD constraint. This reorganizes the original features into new features with a closer distribution. Such a strategy is able to alleviate the incompatibility of different body components while being more conducive to extracting the interaction. We take the alignment process of the major body and left hand as an example.

Cross Neutralization (CN). Given the intra-context $\{\mathbf{S}_l, \mathbf{S}_m\}$, we introduce a learnable factor $\alpha \in [0.5, 1]$ to constitute the fused distribution to neutralize the distributions discrepancy between $\mathbf{S}_l$ and $\mathbf{S}_m$. Formally, the CN$(\cdot)$ is defined as:

$$\begin{aligned}
\boldsymbol{\mu}_{lm,\alpha} &= \alpha\boldsymbol{\mu}_l + (1-\alpha)\boldsymbol{\mu}_m, \boldsymbol{\sigma}_{lm,\alpha} = \alpha\boldsymbol{\sigma}_l + (1-\alpha)\boldsymbol{\sigma}_m, \\
\boldsymbol{\mu}_{ml,\alpha} &= \alpha\boldsymbol{\mu}_m + (1-\alpha)\boldsymbol{\mu}_l, \boldsymbol{\sigma}_{ml,\alpha} = \alpha\boldsymbol{\sigma}_m + (1-\alpha)\boldsymbol{\sigma}_l, \\
\mathbf{s}_l' &= \frac{\mathbf{s}_l - \boldsymbol{\mu}_{lm,\alpha}}{\sqrt{\epsilon + \boldsymbol{\sigma}_{lm,\alpha}^2}}, \mathbf{s}_m' = \frac{\mathbf{s}_m - \boldsymbol{\mu}_{ml,\alpha}}{\sqrt{\epsilon + \boldsymbol{\sigma}_{ml,\alpha}^2}},
\end{aligned} \tag{4}$$

where $\boldsymbol{\mu}_l = \text{Avg}(\mathbf{S}_l)$ and $\boldsymbol{\sigma}_l = \text{Var}(\mathbf{S}_l)$ are the mean and variance of intra-context features $\mathbf{S}_l$; $\boldsymbol{\mu}_m$ and $\boldsymbol{\sigma}_m$ can be obtained similarly; Avg$(\cdot)$ and Var$(\cdot)$ is the operation

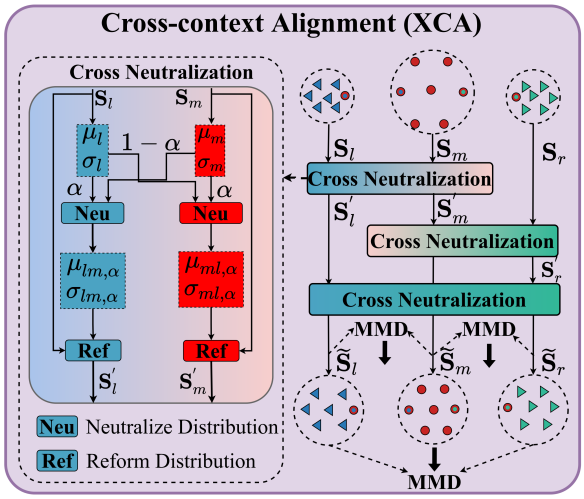

Figure 3: Based on $\{S_l, S_m, S_r\}$, XCA applies circular cross neutralization and discrepancy constraint (MMD) to alleviate the heterogeneity across components and generate the homogeneous features.

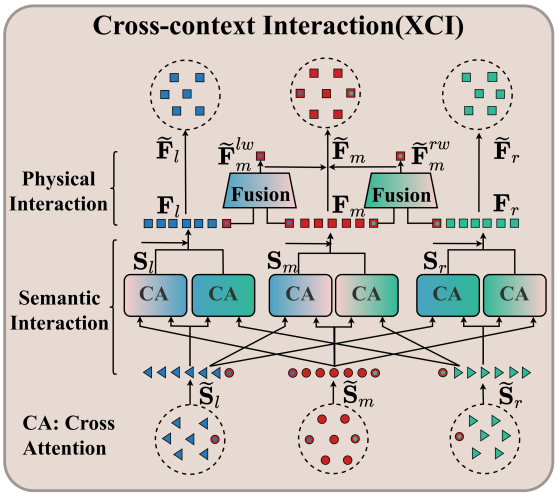

Figure 4: Taking $\{\widetilde{S}_l, \widetilde{S}_m, \widetilde{S}_r\}$ as the input, the XCI explores the pairwise interactivity of different parts from both the semantic and physical interaction within the whole-body.

to calculate mean and variance along the joint dimension; $\boldsymbol{\mu}_l, \boldsymbol{\mu}_m, \boldsymbol{\sigma}_l, \boldsymbol{\sigma}_m \in \mathbb{R}^H$; $\boldsymbol{\mu}_{lm,\alpha}$ ($\boldsymbol{\sigma}_{lm,\alpha}$) $\in \mathbb{R}^H$ are the mean and variance vector of fused features distribution; $\mathbf{s}_m$ ($\mathbf{s}_l$), $\mathbf{s}'_m$ ($\mathbf{s}'_l$) $\in \mathbb{R}^H$ are the row vector of intra-context features $\mathbf{S}_m$ ($\mathbf{S}_l$) and fused features $\mathbf{S}'_m \in \mathbb{R}^{D_m \times H}$ ($\mathbf{S}'_l \in \mathbb{R}^{(D_l+3)\times H}$); $\epsilon = \mathrm{e}^{-5}$ is a factor to avoid numerical issues. To further part-to-part alignment, we extend CN(·) to the circular version:

$$\begin{aligned} \mathbf{S}'_l, \mathbf{S}'_m = \mathrm{CN}(\mathbf{S}_l, \mathbf{S}_m, \alpha),\ \widetilde{\mathbf{S}}_m, \mathbf{S}'_r = \mathrm{CN}(\mathbf{S}'_m, \mathbf{S}_r, \beta), \\ \widetilde{\mathbf{S}}_r, \widetilde{\mathbf{S}}_l = \mathrm{CN}(\mathbf{S}'_r, \mathbf{S}'_l, \gamma), \end{aligned} \tag{5}$$

where $\mathbf{S}'_m$ ($\widetilde{\mathbf{S}}_m$) $\in \mathbb{R}^{D_m \times H}$, $\mathbf{S}'_l$ ($\widetilde{\mathbf{S}}_l$) $\in \mathbb{R}^{(D_l+3)\times H}$, and $\mathbf{S}'_r$ ($\widetilde{\mathbf{S}}_r$) $\in \mathbb{R}^{(D_r+3)\times H}$ are the intermediate (output) features of the circular CN(·); β and γ are the factors similar to α, which are updated in the training phase.

Discrepancy Constraint. We apply maximum mean discrepancy (MMD) to alleviate the part-to-part discrepancy.

$$\mathcal{L}^{dis}_{lm} = \mathrm{MMD}(\mathrm{Avg}(\widetilde{\mathbf{S}}_l), \mathrm{Avg}(\widetilde{\mathbf{S}}_m)), \tag{6}$$

where Avg(·) is the average operation along the spatial dimension, and Avg($\widetilde{\mathbf{S}}_l$)/Avg($\widetilde{\mathbf{S}}_m$)) $\in \mathbb{R}^{1\times H}$. $\mathcal{L}^{dis}_{mr}$ and $\mathcal{L}^{dis}_{rl}$ can be obtained similarly.

With the cross neutralization and discrepancy constraint, where the distribution discrepancy of intra-context features is reduced, our proposed cross-context alignment (XCA) could alleviate the heterogeneity of different intra-context. Next, we present cross-context interaction (XCI) to explore the interactions within the whole body that provides the vital cues to perceive future human intention.

Cross-Context Interaction (XCI)

In contrast to the person↔person external interaction (Guo et al. 2022), the body↔hands/hands↔hands involves the internal interaction across different parts within the whole body. To be precise, it includes both semantic interaction (driven by the collaboration of different parts to perform a specific action) and physical interaction (inherent in the chain link via the body↔hands wrist). Thus, we present a variant of cross-attention (Hao et al. 2017) to capture the semantic and physical interactivity of various human parts.

Semantic Interaction. The relevance of the three body parts is mainly derived from the mutual semantic interaction within the integrated action. For example, for the action of eating, fingers and head joints have strong correlations. Therefore, we aim to model the semantic dependency across components via cross-attention mechanism (Hao et al. 2017). We take the cross-context semantic interaction between the major body and left hand as an example. The whole process is described as:

$$\begin{aligned} &\mathbf{F}^{(1)}_{lm} = \widetilde{\mathbf{S}}_m, \mathbf{Q}^{(n)}_m = \mathbf{F}^{(n)}_{lm}\mathbf{W}^{(n)}_m \\ &\mathbf{K}^{(n)}_l = \widetilde{\mathbf{S}}_l\mathbf{W}^{(n)}_l, \mathbf{V}^{(n)}_l = \widetilde{\mathbf{S}}_l\mathbf{W}^{(n)}_{l'}, \\ &\mathbf{M}^{(n)}_{att,lm} = \mathrm{Softmax}(\mathbf{Q}^{(n)}_m \mathbf{K}^{(n)^{\mathrm{T}}}_l), \\ &\mathbf{F}^{(n+1)}_{lm} = \mathbf{F}^{(n)}_{lm} + \mathrm{FFN}(\mathbf{M}^{(n)}_{att,lm}\mathbf{V}^{(n)}_l), \end{aligned} \tag{7}$$

where $\mathbf{F}^{(n)}_{lm}$ ($\mathbf{F}^{(n+1)}_{lm}$) is the input (output) features; The input feature of the first layer is $\widetilde{\mathbf{S}}_m \in \mathbb{R}^{D_m \times H}$ and the output features of the last layer $\mathbf{F}^{(last)}_{lm} \in \mathbb{R}^{D_m \times H}$; $\mathbf{W}^{(n)}_{l'}$, $\mathbf{W}^{(n)}_l$ and $\mathbf{W}^{(n)}_m$ are the projection matrix with the size of $H \times H$; $\mathbf{Q}^{(n)}_m \in \mathbb{R}^{D_m \times H}$, $\mathbf{K}^{(n)}_l \in \mathbb{R}^{(D_l+3)\times H}$, and $\mathbf{V}^{(n)}_l \in \mathbb{R}^{(D_l+3)\times H}$ are the query, key and value features respectively; $\mathbf{M}^{(n)}_{att,lm} \in \mathbb{R}^{D_m \times (D_l+3)}$ is the attention map calculated by the Softmax(·). FFN(·) is composed of multi-layer perceptrons (MLPs).

For the major body, the semantic relevance with the left hand could be leveraged to fuse the semantic interaction context into itself progressively. Similarly, we can also

obtain the cross-context semantic interaction of the right hand for major body $\mathbf{F}_{rm}^{(last)} \in \mathbb{R}^{D_m \times H}$. Combing the above semantic-related features with the ego features, we can obtain the expressive features $\mathbf{F}_m \in \mathbb{R}^{D_m \times 3H}$:

$$\mathbf{F}_m = \text{Concat}(\mathbf{F}_{rm}^{(last)}; \mathbf{F}_{lm}^{(last)}; \mathbf{S}_m), \tag{8}$$

where Concat$(\cdot)$ is the concatenate operation along the feature dimension. We also attain the features $\mathbf{F}_l \in \mathbb{R}^{(D_l+3)\times 3H}$, $\mathbf{F}_r \in \mathbb{R}^{(D_r+3)\times 3H}$ for left and right hand.

Physical Interaction. As the bridge between the body and hands, the wrist offers direct chain correlation between these two components. Therefore, we apply the 'divide-and-fusion' strategy. That is to say, we first replicate the wrist joint to involve it, as illustrated in section for body and hand independently, and then perform dynamic feature fusion to form the final wrist features. In this way, the physical connection between body parts could be better modeled. Specifically, we identify the feature of the wrist in {major body, left hand} as the complementary pair. It is fed into MLPs to measure the mutual confidence, which are used as a weight to fuse paired features for more informed inference:

$$\widetilde{\mathbf{F}}_m^{lw} = w_{lm}\mathbf{F}_l^{lw} + (1 - w_{lm})\mathbf{F}_m^{lw}, \tag{9}$$

$$w_{lm} = \frac{1}{1 + \exp\left(-\tau * \text{MLP}(\mathbf{F}_l^{lw}, \mathbf{F}_m^{lw})\right)}, \tag{10}$$

where $\mathbf{F}_l^{lw} \in \mathbb{R}^{3\times 3H}$ (or $\mathbf{F}_m^{lw} \in \mathbb{R}^{3\times 3H}$) is the features of the wrist in the left hand (or major body); and w_{lm} is the importance weight. $\widetilde{\mathbf{F}}_m^{lw} \in \mathbb{R}^{3\times 3H}$ is the fused wrist features. τ is a learnable temperature coefficient, jointly trained with all network parameters. Similarly, we obtain the fused wrist features $\widetilde{\mathbf{F}}_m^{rw} \in \mathbb{R}^{3\times 3H}$ for {major body, right hand}.

Then, the final expressive features are further reorganized as follows: **(1)** Removing the left (right) wrist features $\mathbf{F}_l^{lw}/\mathbf{F}_r^{rw}$ from the $\mathbf{F}_l/\mathbf{F}_r$, the final features of left/right hand changes to $\widetilde{\mathbf{F}}_l \in \mathbb{R}^{D_l\times 3H}$ ($\widetilde{\mathbf{F}}_r \in \mathbb{R}^{D_r\times 3H}$); **(2)** As to the body, after physical wrist refinement, the dimension of the feature is unchanged, but with $\mathbf{F}_m^{lw}/\mathbf{F}_m^{rw}$ updated by $\widetilde{\mathbf{F}}_m^{lw}/\widetilde{\mathbf{F}}_m^{rw}$ to generate the final features $\widetilde{\mathbf{F}}_m \in \mathbb{R}^{D_m\times 3H}$; **(3)** The resulting features $\{\widetilde{\mathbf{F}}_l, \widetilde{\mathbf{F}}_m, \widetilde{\mathbf{F}}_r\}$ are then followed by a predictor composed of a MLP and IDCT to regress the final features into predicted sequence $\{\hat{\mathbf{Y}}_m, \hat{\mathbf{Y}}_l, \hat{\mathbf{Y}}_r\}$, where $\hat{\mathbf{Y}}_l \in \mathbb{R}^{D_l\times\Delta T}$, $\hat{\mathbf{Y}}_m \in \mathbb{R}^{D_m\times\Delta T}$, and $\hat{\mathbf{Y}}_r \in \mathbb{R}^{D_r\times\Delta T}$.

Training Loss. Prediction Loss $\mathcal{L}_l^p$ is defined to to measure the accuracy of the predicted 3D coordinates, we calculate the mean per joint position error:

$$\mathcal{L}_l^p = \frac{1}{N_l\Delta T}\sum_{n=1}^{N_l}\sum_{t=1}^{\Delta T}\|\hat{\mathbf{x}}_{n,t} - \mathbf{x}_{n,t}\|, \tag{11}$$

where $\hat{\mathbf{x}}_{n,t} \in \mathbb{R}^3$ denotes the predicted n-th joint position in frame t, $\mathbf{x}_{n,t}$ the corresponding ground truth (GT). N_l the number of joints in the left hand skeleton. Similarly, we can also achieve $\mathcal{L}_r^p$ and $\mathcal{L}_m^p$ for right hand and body, forming the prediction loss of whole body $\mathcal{L}^p = \mathcal{L}_l^p + \mathcal{L}_m^p + \mathcal{L}_r^p$.

Additionally, to further consider the hand semantics, we preprocess the gestures to be aligned with its wrist:

$$\mathcal{L}_l^{pw} = \frac{1}{N_l\Delta T}\sum_{n=1}^{N_l}\sum_{t=1}^{\Delta T}\|\hat{\mathbf{x}}_{n,t}^{w} - \mathbf{x}_{n,t}^{w}\|, \tag{12}$$

where $\hat{\mathbf{x}}_{n,t}^{w} \in \mathbb{R}^3$ denotes the predicted n-th joint position aligned with the left wrist, $\mathbf{x}_{n,t}^{w}$ is the corresponding GT. Then, we can obtain the fine-grained prediction loss of two hands $\mathcal{L}^{pw} = \mathcal{L}_l^{pw} + \mathcal{L}_r^p$.

Since bone length is fixed for a human skeleton, we introduce the bone length loss:

$$\mathcal{L}_l^b = \frac{1}{(N_l-1)\Delta T}\sum_{n=1}^{(N_l-1)}\sum_{t=1}^{\Delta T}|\hat{b}_{n,t} - b_n|, \tag{13}$$

where $\hat{b}_{n,t}$ denotes the length of n-th bone, and b_n the GT. $\mathcal{L}^b = \mathcal{L}_l^b + \mathcal{L}_m^b + \mathcal{L}_r^b$ the bone length loss of whole body.

To alleviate the features heterogeneity of different body components, we utilize the minimum distribution discrepancy error, proposed in section , as the alignment Loss

$$\mathcal{L}^a = \mathcal{L}_{lm}^{dis} + \mathcal{L}_{mr}^{dis} + \mathcal{L}_{rl}^{dis} \tag{14}$$

Final Loss is the weighted sum of the above losses:

$$\mathcal{L} = \lambda_1\mathcal{L}^p + \lambda_2(\mathcal{L}^{pw} + \mathcal{L}^b) + \lambda_3\mathcal{L}^a, \tag{15}$$

where λ_1, λ_2, λ_3 are the trade-off parameters.

Experiments

Datasets. To our knowledge, previous widely-used datasets, *e.g.,* H3.6M (Ionescu et al. 2013), 3DPW (von Marcard et al. 2018), only record the major body motions (without human hands). To be compatible with our proposed novel task, here we select the GRAB (Taheri et al. 2020). It is a recently released dataset with over 1.6 million frames of 10 different actors performing a total of 29 actions. GRAB provides SMPL-X (Pavlakos et al. 2019) parameters from which we extract 25 joints (3D position) defined as the body ($N_m = 25$), and each hand is represented as 15-joints ($N_l = N_r = 15$).

Baselines. We note that for 3D whole-body human motions forecasting, there are no direct methods for comparisons. Therefore, to comprehensively investigate the proposed EAI, we select 4 SOTA approaches of standard major-joint prediction as our baselines, including LTD (Mao et al. 2019), DMGNN (Li et al. 2020), PGBIG (Ma et al. 2022), SPGSN (Li et al. 2022). Notably, all baselines are based on GCNs to consider the N-joint human skeleton ($N = 17$ or $N = 25$). To a fair comparison, we retrain the baselines under the following training setups.

We apply two training strategies to investigate this new task. **(1)** For the **divided (D)** training, we separately train the baselines for each human components. This independent strategy lacks the interaction of components and thus can be used to illustrate the effectiveness of XCI. **(2)** For the **united (D)** training, we extend the node number of GCNs to 55 ($N_m = 25$, $N_l = N_r = 15$), as in our experimental setup. This strategy implicitly contains the cross-context interaction via a whole-body graph but does not consider the heterogeneity of different body parts. Therefore, it is used to demonstrate the effectiveness of XCA.

	Body Parts	Major body			Left Hands			Right Hands			Left Hands (AW)			Right Hands (AW)		
	Time (sec)	0.2	0.4	1.0	0.2	0.4	1.0	0.2	0.4	1.0	0.2	0.4	1.0	0.2	0.4	1.0
Divided	LTD (D)	8.7	18.9	48.7	19.7	57.0	181.5	33.3	77.5	195.6	9.1	18.3	41.4	17.2	28.3	53.1
	DMGNN (D)	11.2	23.1	53.5	24.8	62.0	190.1	38.1	83.0	205.7	10.0	21.7	44.4	21.6	32.6	60.5
	PGBIG (D)	10.4	21.7	52.8	22.8	61.5	186.7	37.6	82.4	203.9	10.5	22.2	43.5	21.5	31.1	58.7
	SPGSN (D)	9.3	21.0	52.6	25.3	61.1	164.2	37.2	81.5	202.8	9.3	18.5	41.6	16.1	28.8	56.9
United	LTD (U)	9.1	19.9	50.2	19.9	50.5	162.5	32.3	74.6	195.5	8.9	17.1	42.5	16.7	29.3	58.3
	DMGNN (U)	13.7	26.4	56.9	22.4	57.3	172.0	36.3	78.9	203.7	9.7	20.3	46.4	19.0	33.2	64.1
	PGBIG (U)	13.2	24.9	54.2	23.0	56.4	165.7	35.0	77.2	199.4	10.2	19.5	45.7	19.1	32.5	62.0
	SPGSN (U)	12.7	24.5	53.4	21.6	55.5	161.6	34.3	75.5	190.8	9.6	18.2	42.3	18.5	31.0	58.2
	EAI (Ours)	**8.3**	**18.7**	**46.8**	**17.7**	**49.2**	**136.4**	**29.8**	**69.0**	**169.0**	**8.6**	**17.3**	**38.8**	**16.2**	**27.8**	**51.6**

Table 1: Average results on all actions with the evaluation metrics MPJPE and MPJPE-AW (in *mm*). (AW), (D) and (U) are the abbreviation of the MPJPE-AW, divided and united training strategies. A lower value means better performance. The best results are highlighted in bold. Notably, EAI is only trained with united training strategies due to the need to explore body components' interactions.

	Action	A1 pass			A2 eat			A3 drink			A4 lift		
	Time (sec)	0.2	0.4	1.0	0.2	0.4	1.0	0.2	0.4	1.0	0.2	0.4	1.0
Major body	LTD (U)	10.2	20.8	42.4	12.1	28.0	**71.7**	12.2	23.4	40.1	7.9	20.3	54.9
	DMGNN (U)	11.7	26.4	40.7	17.9	37.6	88.4	14.5	32.1	58.0	12.1	26.3	61.5
	PGBIG (U)	12.0	26.9	38.9	17.5	36.5	83.2	15.7	30.2	53.2	11.4	24.3	62.4
	SPGSN (U)	13.1	25.8	35.1	18.4	34.8	82.0	15.6	28.9	48.6	10.6	22.6	51.6
	EAI (Ours)	**9.0**	**19.7**	**31.6**	**10.5**	**26.4**	72.5	**10.2**	**19.1**	**30.8**	**6.5**	**16.4**	**44.4**
Left hands	LTD (U)	**24.4**	**52.2**	211.2	21.7	52.5	187.5	51.8	123.4	185.8	21.0	66.3	163.8
	DMGNN (U)	36.7	68.9	196.6	38.6	87.5	234.4	56.2	128.8	265.4	22.2	68.7	181.2
	PGBIG(U)	33.5	66.3	186.2	36.9	88.2	225.6	56.7	126.4	264.6	23.1	66.9	178.4
	SPGSN (U)	30.9	71.1	165.1	36.5	94.6	263.6	51.4	119.8	242.7	20.3	65.2	175.5
	EAI (Ours)	25.4	52.6	**145.1**	**17.8**	**49.0**	**148.7**	**42.5**	**107.7**	**144.4**	**14.3**	**48.4**	**129.7**
Right hands	LTD (U)	37.0	82.1	136.1	35.3	79.3	204.3	22.9	82.2	167.2	25.5	81.5	229.1
	DMGNN (U)	39.2	80.5	129.5	37.5	78.3	215.0	23.5	85.8	221.4	27.3	83.4	231.0
	PGBIG (U)	36.8	78.3	124.6	34.2	76.4	212.5	24.0	87.6	210.5	26.1	82.5	233.7
	SPGSN (U)	33.7	73.0	108.7	31.8	**59.5**	207.6	22.5	92.0	249.4	21.3	76.4	215.6
	EAI (Ours)	**21.7**	**50.3**	**69.6**	**31.8**	70.2	**180.3**	**15.2**	**60.8**	**111.0**	**17.6**	**51.0**	**136.9**

Table 2: Detailed results on common action split with the evaluation metrics MPJPE (in *mm*). (U) is the abbreviation of united training strategy. The best results are highlighted in bold. We observe that for both fine- and coarse-grained motion patterns, our results consistently outperform the competitors. It evidences the compatibility of the EAI for various activities.

Training Details. We employ AdamW (Loshchilov and Hutter 2017) optimizer with an initial learning rate of 0.001 and batch size of 64 to train our model (50 epochs). The learning rate is decayed by 0.96 for every two epochs. The trade-off parameters $\{\lambda_1, \lambda_2, \lambda_3\}$ are set as $\{1, 0.1, 0.001\}$. More details are set in the **supplementary materials**.

Metrics For the whole-body motion, we use the mean per joint position error (MPJPE) (Mao et al. 2019) to measure the 3D prediction accuracy of overall movement. Besides, since there are no baselines for hand prediction, we extend the baseline of major body motion prediction (Mao et al. 2019) into hand prediction and also leverage MPJPE to measure the prediction accuracy. However, the MPJPE of hands is affected by wrist movement severely, which is not able to show subtle hand activities and semantic information. Therefore, we also report the MPJPE-AW after alignment(Martinez et al. 2017) with the wrist.

Comparison With the SOTA Methods

Baselines (U) v.s. Baselines (D). Table 1 shows the average prediction error of all actions between our method and the above four baselines. The baselines are trained with two strategies: divided and united, as illustrated in section . Notably, because the MPJPE of hands is severely affected by wrist movement, we also show the MPJPE-AW on the prediction of delicate hand movement. Compared with the divided strategy, the predicted results for the body are worse when training unitedly. And the hands' results show opposite trends on the two metrics. The above result reveals that: **(1)** The interaction is indeed meaningful to improve prediction accuracy (MPJPE of hands is lower). **(2)** However, the implicit modeling of interaction within a whole-body graph may bring negative mutual interference (MPJPE of body and MPJPE-AW of hands are higher) because major body and gestures have heterogeneous motion patterns.

EAI v.s. Baselines (U&D). Our proposed EAI addresses the above two limitations of existing methods. **(1)** As to the united strategy, EAI is superior to all baselines by a large margin. It verifies the effectiveness of cross-context alignment (XCA), which considers the motion heterogeneity of different body parts. **(2)** Compared with the baseline results using the divided strategy, our method is better, which demonstrates that cross-context interaction (XCI) across body components is vital. Both rough (major joints)

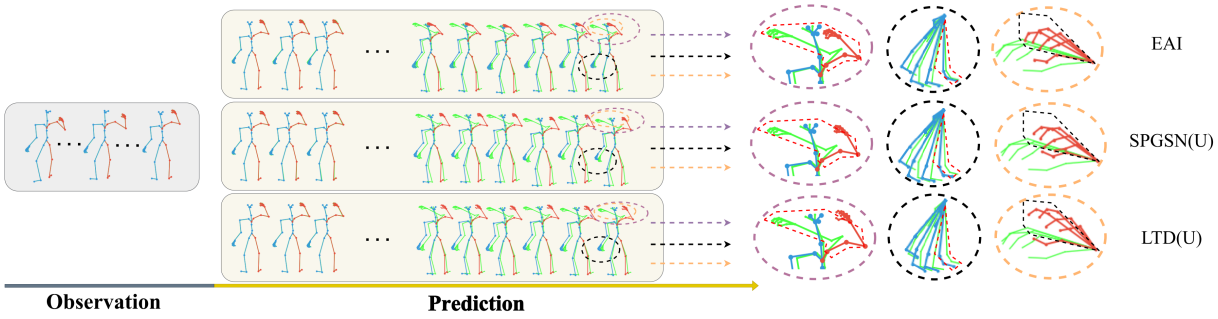

Figure 5: Predicted whole-body poses visualization (skeleton). The past sequence is in a grey box, and the predicted ones are in yellow boxes. The GT and predicted poses are denoted as green and blue/red skeletons, respectively. As highlighted by the dashed ellipse boxes, both performances of fine-grained (body) and coarse-grained (gestures) motion are enhanced.

and delicate (gestures) properties are cross-facilitated to achieve a higher-fidelity prediction via the EAI framework.

Compatibility. Table 2 shows more detailed results on common action with the evaluation metrics MPJPE. The error obtained by our method is smaller than others in most cases. The activity with both fine-grained (drink & eat) and coarse-grained (lift & pass) motion patterns achieve more improvements than the baseline approaches, which evidences the compatibility of our proposed EAI. Moreover, the enhanced performance on all body parts also verifies the necessity of considering both the heterogeneity and interactivity across different body parts. Results of the other actions can be found in the **supplementary material**.

Visualization. In Figure 5, we show the whole-body qualitative results of the 'play' action via the skeletal form. As highlighted by the purple dashed ellipse, the absolute prediction of upper limbs and hands is much closer to ground truth (denoted by solid green lines). It demonstrates that expressive context information extracted from EAI leads to the overall refinement of coarse-fined motion. Besides, the other two dashed ellipses show fine-grained gestures by aligning the hand sequence with the wrist. We observe that EAI still outperforms other baselines in the relative results, which illustrates that the delicate semantic information of gestures could be better considered. The results of fine-grained and coarse-grained motion are enhanced, verifying the significance of co-analyzing different body components for the novel whole-body pose forecasting task.

Ablation Studies

We conduct ablation studies on model architecture for deeper analysis. More discussions are in the **supplementary materials**. We run experiments under the condition of separately removing the XCA and XCI, as well as the following sub-modules: (a) cross neutralization (CN), (b) discrepancy constraint (DC) in XCA; (c) semantic interaction (SI), (d) physical interaction (PI) in XCI.

Table 3 reports the detailed results. (1) Without the CN and DC, the prediction error is 66.6mm, which is a noticeable performance drop, demonstrating the necessity to alleviate distribution discrepancy. Removing CN/DC, the average error increases by 2.5/1.4mm. It shows the CN is more critical in XCA; (2) Excluding the entire XCI, the prediction error drastically increases from 61.9mm to 68.7mm. This gap is larger than the case without the whole XCA, indicating that the interaction extraction is more vital than heterogeneity reduction relatively. Remarkably, the prediction error of XCI (*w/o* SI) / XCI (*w/o* PI) increased by 5.6/2.6mm. It reveals that the semantic relevance of body components is more valuable to perceive motion properties.

CN	DC	PI	SI	0.2s	0.4s	1.0s	Avg.
	✓	✓	✓	**16.7**	**40.7**	90.4	64.4
✓		✓	✓	17.0	41.3	87.9	63.3
		✓	✓	17.0	42.8	93.7	66.6
✓	✓		✓	**16.7**	41.1	89.8	64.5
✓	✓	✓		17.0	42.5	94.2	67.5
✓	✓			17.8	43.2	95.0	68.7
✓	✓	✓	✓	**16.7**	**40.7**	**85.8**	**61.9**

Table 3: Ablation experiments of model architecture. The best (second-best) result is highlighted in bold (underlined).

Conclusion

In this work, we introduce a new task: expressive forecasting of 3D whole-body human motions. To tackle this challenge, we propose a novel Encoding-Alignment-Interaction (EAI) framework that takes into account the heterogeneous information within the whole body and the collaboration among various human components. Our approach jointly considers the heterogeneous information within the whole body and the interaction among various human components. Compared with conventional predictive algorithms, EAI could cross-facilitate both coarse- (body) and fine-grained (gestures) properties. Experiments demonstrate that the proposed approach achieve the superior performance and outperforms the state-of-the-art methods. Considering the downstream application of whole-body forecasting, we conclude that the proposed model is of practical importance.

Acknowledgments

This work was supported by the National Science and Technology Innovation 2030 - Major Project (Grant No. 2022ZD0208800), and NSFC General Program (Grant No. 62176215). This work was supported in part by the National Natural Science Foundation of China (62306141), in part by the Jiangsu Funding Program for Excellent Postdoctoral Talent (2022ZB269), in part by the Natural Science Foundation of Jiangsu Province (BK20220939), and in part by the China Postdoctoral Science Foundation (2022M721629).

References

Adeli, V.; Ehsanpour, M.; Reid, I.; Niebles, J. C.; Savarese, S.; Adeli, E.; and Rezatofighi, H. 2021. TRiPOD: Human Trajectory and Pose Dynamics Forecasting in the Wild. In *IEEE/CVF International Conference on Computer Vision*, 13390–13400.

Butepage, J.; Black, M. J.; Kragic, D.; and Kjellstrom, H. 2017. Deep representation learning for human motion prediction and classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6158–6166.

Bütepage, J.; Kjellström, H.; and Kragic, D. 2018. Anticipating many futures: Online human motion prediction and generation for human-robot interaction. In *2018 IEEE/CVF International Conference on Robotics and Automation (ICRA)*, 4563–4570.

Cai, Y.; Huang, L.; Wang, Y.; Cham, T.-J.; Cai, J.; Yuan, J.; Liu, J.; Yang, X.; Zhu, Y.; Shen, X.; et al. 2020. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*, 226–242.

Cai, Y.; Wang, Y.; Zhu, Y.; Cham, T.-J.; Cai, J.; Yuan, J.; Liu, J.; Zheng, C.; Yan, S.; Ding, H.; et al. 2021. A Unified 3D Human Motion Synthesis Model via Conditional Variational Auto-Encoder. In *IEEE/CVF International Conference on Computer Vision*, 11645–11655.

Corona, E.; Pumarola, A.; Alenya, G.; and Moreno-Noguer, F. 2020. Context-aware human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6992–7001.

Cui, Q.; and Sun, H. 2021. Towards Accurate 3D Human Motion Prediction From Incomplete Observations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4801–4810.

Dang, L.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2021. MSR-GCN: Multi-Scale Residual Graph Convolution Networks for Human Motion Prediction. In *IEEE/CVF International Conference on Computer Vision*, 11467–11476.

Diller, C.; Funkhouser, T.; and Dai, A. 2022. Forecasting characteristic 3D poses of human actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15914–15923.

Ding, P.; and Yin, J. 2021. Uncertainty-aware Human Motion Prediction. *arXiv preprint arXiv:2107.03575*.

Ding, P.; and Yin, J. 2022. Towards more realistic human motion prediction with attention to motion coordination. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9): 5846–5858.

Feng, Y.; Choutas, V.; Bolkart, T.; Tzionas, D.; and Black, M. J. 2021. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, 792–804.

Gui, L.-Y.; Wang, Y.-X.; Liang, X.; and Moura, J. M. 2018. Adversarial geometry-aware human motion prediction. In *European Conference on Computer Vision*, 786–803.

Guo, W.; Bie, X.; Alameda-Pineda, X.; and Moreno-Noguer, F. 2022. Multi-Person Extreme Motion Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13053–13064.

Hao, Y.; Zhang, Y.; Liu, K.; He, S.; Liu, Z.; Wu, H.; and Zhao, J. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 221–231.

Hidalgo, G.; Raaj, Y.; Idrees, H.; Xiang, D.; Joo, H.; Simon, T.; and Sheikh, Y. 2019. Single-network whole-body pose estimation. In *IEEE/CVF International Conference on Computer Vision*, 6982–6991.

Honda, Y.; Kawakami, R.; and Naemura, T. 2020. RNN-based Motion Prediction in Competitive Fencing Considering Interaction between Players. In *BMVC*.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. 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.

Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; and Luo, P. 2020. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, 196–214.

Li, J.; Yang, F.; Ma, H.; Malla, S.; Tomizuka, M.; and Choi, C. 2021. RAIN: Reinforced Hybrid Attention Inference Network for Motion Forecasting. In *IEEE/CVF International Conference on Computer Vision*.

Li, M.; Chen, S.; Zhang, Z.; Xie, L.; Tian, Q.; and Zhang, Y. 2022. Skeleton-Parted Graph Scattering Networks for 3D Human Motion Prediction. In *European Conference on Computer Vision*.

Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; and Tian, Q. 2020. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 214–223.

Li, R.; Wang, S.; Zhu, F.; and Huang, J. 2018. Adaptive graph convolutional neural networks. In *AAAI Conference on Artificial Intelligence*, 1.

Liu, X.; Yin, J.; Liu, J.; Ding, P.; Liu, J.; and Liu, H. 2020. Trajectorycnn: a new spatio-temporal feature learning network for human motion prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6): 2133–2146.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, T.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2022. Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6437–6446.

Mao, W.; Liu, M.; and Salzmann, M. 2020. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, 474–489.

Mao, W.; Liu, M.; Salzmann, M.; and Li, H. 2019. Learning trajectory dependencies for human motion prediction. In *International Conference on Computer Vision*, 9489–9497.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2640–2649.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10975–10985.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-Conditioned 3D Human Motion Synthesis With Transformer VAE. In *IEEE/CVF International Conference on Computer Vision*, 10985–10995.

Rong, Y.; Shiratori, T.; and Joo, H. 2021. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE/CVF International Conference on Computer Vision*, 1749–1759.

Ruiz, A. H.; Gall, J.; and Moreno-Noguer, F. 2018. Human Motion Prediction via Spatio-Temporal Inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7134–7143.

Taheri, O.; Ghorbani, N.; Black, M. J.; and Tzionas, D. 2020. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision*, 581–600.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*.

von Marcard, T.; Henschel, R.; Black, M.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In *European Conference on Computer Vision*.

Wang, J.; Xu, H.; Narasimhan, M.; and Wang, X. 2021. Multi-Person 3D Motion Prediction with Multi-Range Transformers. *Advances in Neural Information Processing Systems*, 34: 6036–6049.

Yuan, Y.; and Kitani, K. 2020. Dlow: Diversifying Latent Flows for Diverse Human Motion Prediction. In *European Conference on Computer Vision*, 346–364.

Zhang, Y.; Black, M. J.; and Tang, S. 2021. We Are More Than Our Joints: Predicting How 3D Bodies Move. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3372–3382.

Zhong, C.; Hu, L.; Zhang, Z.; Ye, Y.; and Xia, S. 2022. Spatio-Temporal Gating-Adjacency GCN for Human Motion Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6447–6456.