Forecasting of 3D Whole-body Human Poses with Grasping Objects

Haitao Yan[1] Qiongjie Cui[2,*] Jiexin Xie[1] Shijie Guo[1]
[1]Academy for Engineering & Technology, Fudan University, Shanghai, China
[2]Nanjing University of Science and Technology, Nanjing, China
htyan20@fudan.edu.cn, cuiqiongjie@126.com

Abstract

In the context of computer vision and human-robot interaction, forecasting 3D human poses is crucial for understanding human behavior and enhancing the predictive capabilities of intelligent systems. While existing methods have made significant progress, they often focus on predicting major body joints, overlooking fine-grained gestures and their interaction with objects. Human hand movements, particularly during object interactions, play a pivotal role and provide more precise expressions of human poses. This work fills this gap and introduces a novel paradigm: forecasting 3D whole-body human poses with a focus on grasping objects. This task involves predicting activities across all joints in the body and hands, encompassing the complexities of internal heterogeneity and external interactivity. To tackle these challenges, we also propose a novel approach: C^{3}HOST, cross-context cross-modal consolidation for 3D whole-body pose forecasting, effectively handles the complexities of internal heterogeneity and external interactivity. C^{3}HOST involves distinct steps, including the heterogeneous content encoding and alignment, and cross-modal feature learning and interaction. These enable us to predict activities across all body and hand joints, ensuring high-precision whole-body human pose prediction, even during object grasping. Extensive experiments on two benchmarks demonstrate that our model significantly enhances the accuracy of whole-body human motion prediction. The project page is available at https://sites.google.com/view/c3host.

1. Introduction

Forecasting upcoming 3D human poses, conditioned on the historical ones, is an essential task in human-robot interaction (HRI) [5, 6, 21, 22, 25, 30, 33, 34, 39, 50, 62, 65].

Although this attractive domain has achieved commendable success, it remains a major limitation, *i.e.*, existing methods concentrate on forecast major human body (17 or 25 joints), neglecting the fine-grained gestures and their interaction with objects [24, 33, 34, 43, 46]. For person-to-object interaction, hand activities present the pivotal significance, especially in the context of activities involving object grasping, *e.g.*, lifting a water bottle. Furthermore, we note that human hand, in combination with major body activities, is a more precise expressions of human poses, reflecting the behaviors of grasping/manipulating objects, instructions for robots, and human underlying intentions [4, 13, 35, 40, 64].

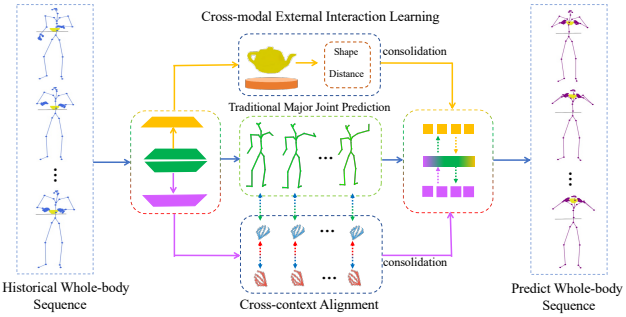

Figure 1. Previous studies focus solely on predicting the motion of the main joints of body, neglecting hand actions and environmental factors. To address this gap, this paper introduces a new task: the forecasting of 3D whole-body human poses with grasping objects.

This work aims to address this meaningful issue, simultaneous predicting activities across all joints in both body and hands, with a particular emphasis on human grasping of objects, as shown in Figure 1. For this novel task, the following aspects need to be solved:

- *Internal Heterogeneity:* Gestures are governed by the physical structure of human body; but meanwhile, these elements involve the heterogeneous motion patterns, including magnitude, motion dynamics, *etc*.
- *External Interactivity:* The modalities of human motion and object are distinct, and the human-object interaction undergoes dynamic fluctuations across the various timestamps and human components (sometimes left hand, sometimes right hand). This intricacies gives rise to varying physical interactions.

*Corresponding author

To address these challenges, we introduce a novel method, C^{3}HOST, cross-context cross-modal consolidation for 3D whole-body pose forecasting with grasping objects. It comprises three main steps: heterogeneous internal motion context encoding, cross-context alignment, and cross-modal external interaction learning. Considering the diverse motion contexts, we initially partition the entire human body into three components: body, left hand, and right hand. This allows us to extract separate spatiotemporal correlations to mitigate mutual interference. The rationale behind this lies in the fact that the movement patterns of the left and right hands interact and influence each other, particularly during object interactions, where the interaction intensity between both hands and the object varies. We therefore employ the maximum mean discrepancy [9, 18] for aligning and smoothing the features of these components to eliminate feature heterogeneity. We also introduce a circular cross-attention structure to capture correlations between different body parts during motion [10, 37, 42]. Concerning external human-object features, we extract spatial structural features of objects and align them with the skeletal features of the whole-body human in the same spatial context. Finally, we incorporate a graph attention network (GAT) to aggregate information from four different modalities and contexts (object, body, two hands) [55, 57]. It facilitates the interaction between various modalities while avoiding information redundancy resulting from direct feature fusion. Moreover, the attention mechanism is employed to differentiate the strength of correlations between different modal features. Additionally, we integrate gated distance information to capture the evolving impact of objects on the major body and both hands over time. It is worth noting that the proposed C^{3}HOST framework can simultaneously consider the interaction of heterogeneous motion contexts within the internal body and the interaction of external human-object modal features, thereby achieving high-precision whole-body pose prediction when grasping objects.

Our main contributions are: 1) We propose to solve the task of forecasting 3D whole-body human motions, encompassing object interactions and involving the entire human body and hands. 2) To address feature interactions within the body-hand context and human-object interactions, we propose two novel approaches: heterogeneous internal motion context encoding/alignment and cross-modal external interactivity learning. 3) Extensive experiments demonstrate that our model significantly enhances the accuracy of whole-body motion prediction.

2. Related Work

Human Motion Forecasting. Recent research has witnessed a surge of innovative methods tackling human motion prediction from diverse perspectives [7, 11, 12, 19, 29, 33, 46, 50, 59, 61]. Among these, GCNs have recently garnered considerable attention [14, 15, 35, 36, 41, 44]. [44] employs GCNs to encode spatial information, thereby enhancing the extraction of joint spatial features and proposing a straightforward 3D pose prediction network. [15, 36] explore the abstraction of spatial features from human body joints, yielding posture features at various scales. Multi-scale residual graph convolution network [15], comprising an ensemble of GCNs, integrates and decodes these multi-scale features. Further contributions [12, 13, 64] have extended the application to real-world scenarios. [13] addresses the occlusion issue in motion prediction, while [12], for the first time, considers leveraging environmental feature to constrain motion prediction. [64] incorporate scene understanding and human gaze as distinct modalities, applying them as prior knowledge to enforce constraints related to the physical environment.

Despite these advancements, the primary focus has been on predicting body motion, with scant attention to the nuanced analysis of human gestures and object interactions. Given the practical HRI implications, achieving precise whole-body motion prediction is a critical research goal. Our work recognizes this gap and aspires to offer a comprehensive solution.

Cross-modal Feature Learning. Multimodal feature integration is crucial for a holistic comprehension of the impact of various object modalities on human motion. Transformer encoders have demonstrated efficacy in fusing multimodal features [31, 49, 54]. In a parallel vein, [38] employs maximum mean discrepancy to adjust the statistical distribution of dual modality data, thereby aligning the distributions and enhancing the fusion of multimodal information [9, 18]. Nevertheless, multimodal fusion in the context of human-object interaction encounters distinct challenges. While object modalities remain static across motion sequences, the intensity of interaction between human motion modalities and object modalities fluctuates dynamically. To tackle this, we introduce a cutting-edge technique termed cross-modal external interaction learning. It advocates a multimodal fusion strategy predicated on graph attention networks, leveraging weight and distance optimizations to catalyze inter-modality feature amalgamation, thus optimizing the correlation and complementarity between heterogeneous modalities.

Contextual Interaction. The human body's movement is intrinsically connected with its surroundings, characterized by ongoing interactions with the environment and objects. By integrating contextual interactions, including environmental cues and object-related factors, we can constrain human pose, position, and trajectory [1, 2, 27, 58, 63]. In motion generation, environmental context is commonly employed as a conditional signal. Works such as [20, 52, 60] harness diverse inputs like environmental context and motion directives to synthesize credible, virtual, and varied

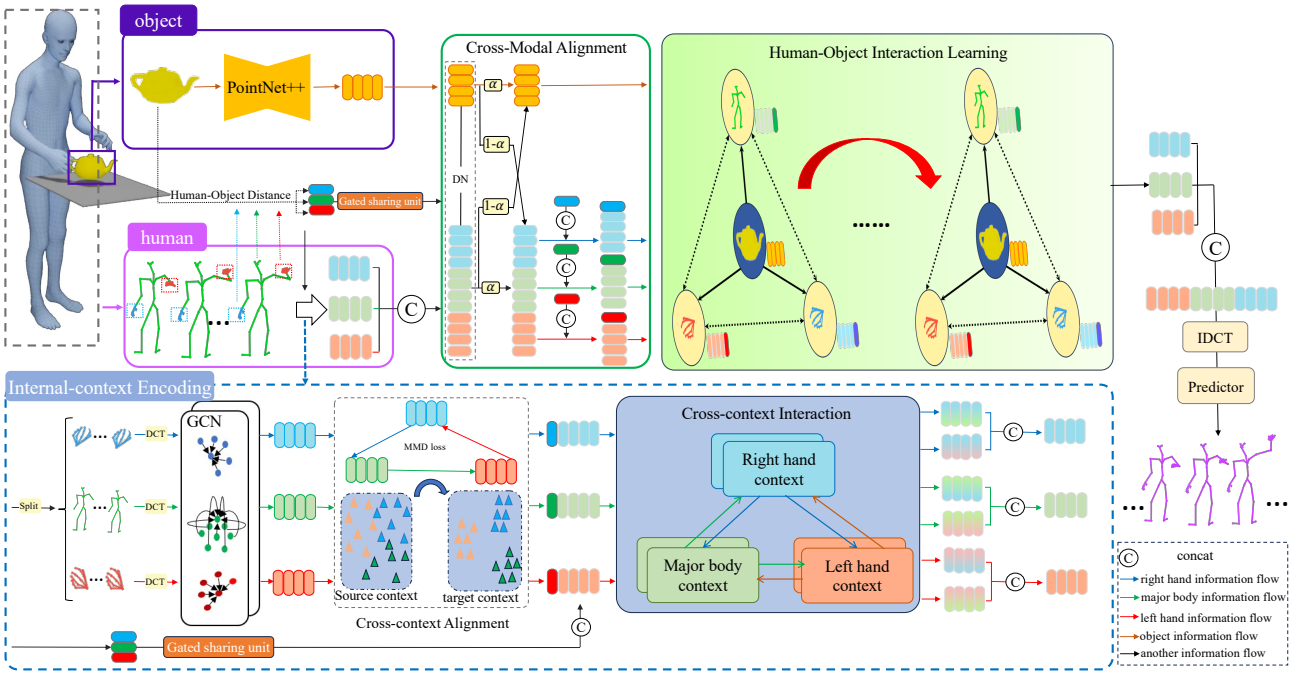

Figure 2. Overall framework of C^{3}HOST. Starting with the input of whole-body sequences $\boldsymbol{X}$ and object mesh $\boldsymbol{S}$, we independently extract spatiotemporal feature of the body, left hand and right hand $\{\boldsymbol{X}_l, \boldsymbol{X}_b, \boldsymbol{X}_r\}$. Considering the heterogeneity and multimodality of data, we introduce two key components: internal-context encoding and cross-modal external interaction learning. The former mines the cross-context features to yield homogeneous features $\{\tilde{\boldsymbol{X}}'_l, \tilde{\boldsymbol{X}}'_b, \tilde{\boldsymbol{X}}'_r\}$, while the latter aims to learn cross-modal external interaction between human and object. Moreover, we also calculate the distance information between the human and object, aligning object features with human joint features in the temporal domain. This alignment is crucial for capturing the nuanced dynamics of human-object interactions. Finally, the obtained features $\{\hat{\boldsymbol{X}}_l, \hat{\boldsymbol{X}}_b, \hat{\boldsymbol{X}}_r\}$ are decoded to predict future whole-body sequences $\{\hat{\boldsymbol{Y}}_{l,b,r}\}$.

full-body human motions. Several researches [12, 64] have also highlighted the benefits of incorporating contextual interactions. For example, [23, 56] concentrate on interpersonal interactions and limb coordination. However, these methods often prioritize forecasting the movements of major body joints, frequently neglecting the significance of intricate hand movements within these interactions. To capture contextual interactions between humans and objects with greater precision, our methodology distills interaction information among various human limb segments, ensuring synchronized limb movements. Concurrently, we incorporate object-specific attributes, like shape, size, and position, to set constraints on hand movements, which enables the precise predictions of 3D whole-body human poses within sophisticated interactive settings.

3. Proposed Method

3.1. Problem Setup

Let $\boldsymbol{X}_{1:T} = [\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_T]$ be a historical human motion sequence with T poses, $\boldsymbol{Y}_{T+1,T+\Delta T} = [\boldsymbol{p}_{T+1}, \boldsymbol{p}_{T+2}, ..., \boldsymbol{p}_{T+\Delta T}]$ be the future poses of length ΔT, human pose forecasting is to learn a mapping function $\mathcal{F} : \boldsymbol{X}_{1:T} \rightarrow \boldsymbol{Y}_{T+1,T+\Delta T}$. In the following, for simplicity, $\boldsymbol{X}, \boldsymbol{Y}$ are used to represent $\boldsymbol{X}_{1:T}$, $\boldsymbol{Y}_{T+1,T+\Delta T}$.

Standard methods typically consider the major joint without the fine-grained hand gestures [14, 36, 44]. Our work expands it to encompass a unified prediction of whole-body motion, including left hand, the body, and right hand. To simplify, we use variables l, b, and r to represent these three parts. In addition, we also aim to learn the human-object grasping, marked as $\boldsymbol{S} \in \mathbb{R}^{3O}$, where O is the number of vertices of the object mesh. Informally, this task can be defined as learning a new mapping $\mathcal{F}_{\text{new}}$:

$$\mathcal{F}_{\text{new}} : \{\boldsymbol{X}_{l,b,r}, \boldsymbol{S}\} \rightarrow \{\boldsymbol{Y}_{l,b,r}\}, \tag{1}$$

where $\boldsymbol{X}_l \in \mathbb{R}^{3N_l \times T}$, $\boldsymbol{X}_b \in \mathbb{R}^{3N_b \times T}$, $\boldsymbol{X}_r \in \mathbb{R}^{3N_r \times T}$, and N_l, N_b, N_r are the number of joints.

3.2. Method Overview

Figure 2 illustrates the overall pipeline of the proposed C^{3}HOST. Given the whole-body motion sequences $\{\boldsymbol{X}_l, \boldsymbol{X}_b, \boldsymbol{X}_r\}$ and object mesh $\boldsymbol{S}$, we first employ DCT [3] and GCN [44] to extract separate spatiotemporal information of the major body, left hand and right hand $\{\boldsymbol{X}'_l, \boldsymbol{X}'_b, \boldsymbol{X}'_r\}$, and use the pre-trained PointNet++ [48] to encode the object $\boldsymbol{S}'$. Due to the heterogeneity of human internal-contexts, we propose a cross-context alignment based on maximum mean discrepancy to obtain homogeneous features for the body and hands. Then, we propose

a circular cross-attention structure to capture internal correlations between different body parts during motion. We also suggest that the distance between the human and object is a substitute for capturing the external human-object interaction. To this end, we calculate the distance information $\{\boldsymbol{d}_{l\leftrightarrow o}, \boldsymbol{d}_{b\leftrightarrow o}, \boldsymbol{d}_{r\leftrightarrow o}\}$ between different human parts and object, and propose a gated sharing unit to decide the importance of distance information to the human-object interaction. Meanwhile, the distribution normalization is used to align the object features and human joint features in the temporal domain. Then, we utilize the graph attention network to aggregate information from 4 different modalities and contexts (object, body, two hands), and to learn the human-object interaction. Finally, the obtained expressive features are decoded by inverse discrete cosine transform (IDCT) to predict future whole-body sequences $\{\hat{\boldsymbol{Y}}_{l,b,r}\}$.

3.3. Internal-context Encoding

Heterogeneous contexts extraction. For different human parts, hand joints exhibits significant variations with limited motion ranges, and distinct motion patterns exist between the two hands and the body. Therefore, we consider the joints of the body, left hand, and right hand as diverse motion contexts, and separately extract the spatiotemporal information. Concretely, based on prior research [24, 34, 44], discrete cosine transform (DCT) is first used to encode temporal correlations of motion sequences:

$$\boldsymbol{X}'_l = \boldsymbol{X}_l \boldsymbol{C},\ \ \boldsymbol{X}'_b = \boldsymbol{X}_b \boldsymbol{C},\ \ \boldsymbol{X}'_r = \boldsymbol{X}_r \boldsymbol{C}, \tag{2}$$

where $\boldsymbol{C}$ is the DCT matrix, and $\boldsymbol{X}'_l$, $\boldsymbol{X}'_b$, $\boldsymbol{X}'_r$ denote the transformed sequence from DCT.

Then, we employ the fully-connected GCN as in [44] to individually encode their spatial features $\boldsymbol{X}''_l$, $\boldsymbol{X}''_b$, $\boldsymbol{X}''_r$.

Cross-context alignment (CCA). To enable information exchange between these diverse context, it is essential to align the information from each of them. Distribution disparities between different motion context manifest in misalignment of elements, including variations in joint density, joint motion range, and degrees of freedom in different body parts [33]. For this issue, we utilize a loss function based on maximum mean discrepancy (MMD) [51] to minimize the distance in feature space between average embeddings:

$$\begin{aligned}
\text{MMD}(\boldsymbol{X}''_l, \boldsymbol{X}''_b) &= \left\| \frac{1}{N_l}\sum_{i=1}^{N_l} \phi(\boldsymbol{X}''_{l,i}) - \frac{1}{N_b}\sum_{j=1}^{N_b} \phi(\boldsymbol{X}''_{b,j}) \right\|^2_{\mathcal{H}}, \\
\text{MMD}(\boldsymbol{X}''_b, \boldsymbol{X}''_r) &= \left\| \frac{1}{N_b}\sum_{i=1}^{N_b} \phi(\boldsymbol{X}''_{b,i}) - \frac{1}{N_r}\sum_{j=1}^{N_r} \phi(\boldsymbol{X}''_{r,j}) \right\|^2_{\mathcal{H}}, \\
\text{MMD}(\boldsymbol{X}''_r, \boldsymbol{X}''_l) &= \left\| \frac{1}{n_r}\sum_{i=1}^{N_r} \phi(\boldsymbol{X}''_{r,i}) - \frac{1}{N_l}\sum_{j=1}^{N_l} \phi(\boldsymbol{X}''_{l,j}) \right\|^2_{\mathcal{H}},
\end{aligned} \tag{3}$$

where $\phi(\cdot)$ is a Gaussian kernel function that fit any distribution into the Hilbert space $\mathcal{H}$. We use $\phi(\cdot)$ to map two joint distributions into high-dimensional space, and then calculate the mathematic expectation between the two distributions to obtain the maximum mean discrepancy.

Cross-context interaction (CCI). Within the human body, a person's torso, head, and limbs engage in internal interactions (*e.g.*, eating). Particularly for grasping movements, these interactions involve semantic interactions between the two hands (cooperative execution of grasping actions by both hands) and constraint interactions between the torso and hands (the physical connection of the torso to the hands). Moreover, the intensity of these internal interactions varies over time and with changes in the distance between the person and the object. Therefore, we propose a circular cross-attention mechanism [26] to model these internal interactions among body components. Taking the homogeneous context interaction between the left and right hands as an example, the process is as follows:

$$\tilde{\boldsymbol{X}}'_l = \boldsymbol{X}''_l + \text{softmax}\left(\frac{\boldsymbol{X}'_l \boldsymbol{W}_q \cdot \boldsymbol{X}'_r \boldsymbol{W}_k}{\sqrt{d}}\right) \boldsymbol{X}'_l \boldsymbol{W}_v, \tag{4}$$

where $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{W}_v$ are learnable weights of cross-attention, d is the dimension. Similarly, we can obtain $\tilde{\boldsymbol{X}}'_b$ and $\tilde{\boldsymbol{X}}'_r$ based on the above method, respectively.

3.4. Cross-modal External Interaction Learning

Object feature extraction. To enhance the features of the hand with object information, this work employs PointNet++ [48] to extract shape features of objects and map them to a higher-dimensional space, denoted as $\boldsymbol{S}'$. To establish temporal correspondence between objects and human joints, we leverage the human-object distance to modulate the variations in internal interaction information that accompany changes in distance. This acts as a surrogate for temporal correlations between humans and objects. In essence, time affects the distance between humans and objects, and this distance, in turn, influences the strength of the interaction. To more accurately utilize human-object distance information, we refine the set $\{\boldsymbol{d}_{b\leftrightarrow o}, \boldsymbol{d}_{l\leftrightarrow o}, \boldsymbol{d}_{r\leftrightarrow o}\}$, which represents the minimum distance between the joints of various human body parts and the object at each time step. We introduce a gated sharing unit to capture the dynamic influence of objects on the major body and both hands over time, defined as follows:

$$\begin{aligned}
\boldsymbol{d}'_{l\leftrightarrow o} &= \sigma(\boldsymbol{d}_{l\leftrightarrow o}\boldsymbol{W}_l + \boldsymbol{b}_l) \otimes \delta(\boldsymbol{d}_{l\leftrightarrow o}\boldsymbol{U}_l + \boldsymbol{c}_l), \\
\boldsymbol{d}'_{b\leftrightarrow o} &= \sigma(\boldsymbol{d}_{b\leftrightarrow o}\boldsymbol{W}_b + \boldsymbol{b}_b) \otimes \delta(\boldsymbol{d}_{b\leftrightarrow o}\boldsymbol{U}_b + \boldsymbol{c}_b), \\
\boldsymbol{d}'_{r\leftrightarrow o} &= \sigma(\boldsymbol{d}_{r\leftrightarrow o}\boldsymbol{W}_r + \boldsymbol{b}_r) \otimes \delta(\boldsymbol{d}_{r\leftrightarrow o}\boldsymbol{U}_r + \boldsymbol{c}_r),
\end{aligned} \tag{5}$$

where σ is the sigmoid function, δ is LeakyReLU with a slope of 0.2. $\boldsymbol{W}$, $\boldsymbol{U}$, $\boldsymbol{b}$, $\boldsymbol{c}$ are the learnable weight and bias, respectively, $\otimes$ represents element-wise product. After adding distance information along the joint coordinate dimension, the human posture features are represented as:

$$\begin{aligned}\tilde{\boldsymbol{X}}_l^{''} &= \text{concat}(\boldsymbol{d}_{l\leftrightarrow o}^{'},\ \tilde{\boldsymbol{X}}_l^{'}),\\ \tilde{\boldsymbol{X}}_b^{''} &= \text{concat}(\boldsymbol{d}_{b\leftrightarrow o}^{'},\ \tilde{\boldsymbol{X}}_b^{'}),\\ \tilde{\boldsymbol{X}}_r^{''} &= \text{concat}(\boldsymbol{d}_{r\leftrightarrow o}^{'},\ \tilde{\boldsymbol{X}}_r^{'}).\end{aligned} \tag{6}$$

We note that $\boldsymbol{d}_{l\leftrightarrow o}^{'}, \boldsymbol{d}_{b\leftrightarrow o}^{'}, \boldsymbol{d}_{r\leftrightarrow o}^{'} \in \mathbb{R}^{1\times T}$, $\tilde{\boldsymbol{X}}_l^{''} \in \mathbb{R}^{(3N_l+1)\times T}$, $\tilde{\boldsymbol{X}}_b^{''} \in \mathbb{R}^{(3N_l+1)\times T}$ and $\tilde{\boldsymbol{X}}_r^{''} \in \mathbb{R}^{(3N_l+1)\times T}$.

Cross-modal alignment (CMA). Objects exert a substantial enlightening influence on human motion. However, a discrepancy often arises in the marginal distributions of human and object modal features post feature mapping, causing a separation that belies their inherent relevance [45]. To address this issue, we incorporate a learnable factor α within the range $[0.5, 1]$ to modulate the distribution shift for elements exhibiting high-confidence correlations. This adjustment strategy harmonizes the mean and variance of multi-modal feature distributions within an unsupervised framework, striving to render the learned source and target representations maximally similar. Because the human body and objects are in the same feature space, we aggregate the features of the three parts of the human body, represented as $\tilde{\boldsymbol{X}}^{''} = \{\tilde{\boldsymbol{X}}_l^{''}, \tilde{\boldsymbol{X}}_b^{''}, \tilde{\boldsymbol{X}}_r^{''}\}$.

The process of aligning the feature of the human joints and object can be achieved. Assume $\tilde{\boldsymbol{Q}}_p$ and $\tilde{\boldsymbol{S}}_o$ are the distribution of human pose feature $\tilde{\boldsymbol{X}}^{''}$ and object shape $\boldsymbol{S}'$, we use the factor α to mix context distributions: $\tilde{\boldsymbol{Q}}_{po} = \alpha\tilde{\boldsymbol{Q}}_p + (1-\alpha)\tilde{\boldsymbol{Q}}_o$, similarly, $\tilde{\boldsymbol{Q}}_{op} = \alpha\tilde{\boldsymbol{Q}}_o + (1-\alpha)\tilde{\boldsymbol{Q}}_p$. Then, we calculate the mean and variance of the mixed distribution: $\mu_{po,\alpha} = Avg(\tilde{\boldsymbol{Q}}_{po})$, $\sigma^2_{op,\alpha} = Var(\tilde{\boldsymbol{Q}}_{op})$, and $\mu_{po,\alpha} = Avg(\tilde{\boldsymbol{Q}}_{po})$, $\sigma^2_{op,\alpha} = Var(\tilde{\boldsymbol{Q}}_{op})$. The aligned distribution can be represented as:

$$\tilde{\boldsymbol{Q}}_p = \frac{\tilde{\boldsymbol{Q}}_p - \mu_{po,\alpha}}{\sqrt{\epsilon + \sigma^2_{po,\alpha}}}, \quad \tilde{\boldsymbol{Q}}_o = \frac{\tilde{\boldsymbol{Q}}_o - \mu_{op,\alpha}}{\sqrt{\epsilon + \sigma^2_{op,\alpha}}}, \tag{7}$$

where $\epsilon = \text{e}^{-5}$ is a small number to avoid numerical issues in case of zero variance. $\tilde{\boldsymbol{Q}}_p$ and $\tilde{\boldsymbol{Q}}_o$ are the aligned feature distribution of human pose and object.

Human-object interaction learning (HOIL). In the context of grasping motions, a strong interaction exists between objects and the hands. However, in contrast to human skeletal features, object features exhibit a clear granularity difference. Furthermore, in the temporal domain, human skeletal features represent time-varying sequences, whereas object features remain constant. Dramatically different modalities often exhibit pronounced feature redundancy, making direct fusion a challenging endeavor for enhancing feature representations. Consequently, we propose a multi-modal fusion method based on the graph attention network [55]. In this approach, the major body, left hand, right hand, and distance are treated as graph nodes, with attention weights employed to discern the significance and relevance of neighboring nodes during the aggregation of information. This indirect information interaction effectively eliminates redundant features between modalities, thereby facilitating the acquisition of improved feature representations. Taking homogeneous human internal-context $\{\tilde{\boldsymbol{X}}_l^{''}, \tilde{\boldsymbol{X}}_b^{''}, \tilde{\boldsymbol{X}}_r^{''}\}$ and human-object external modal features $\tilde{\boldsymbol{S}}_o$ as the input, the whole process is described as:

$$\alpha_{i,j} = \text{softmax}\left(\frac{\exp(\delta(V_i\boldsymbol{h}_i, V_j\boldsymbol{h}_j))}{\sum_{k\in\mathcal{N}_i}\exp(\delta(V_j\boldsymbol{h}_i, V_j\boldsymbol{h}_k))}\right), \tag{8}$$

where $\alpha_{i,j}$ is the weight coefficient of attention, δ is LeakyReLU with a slope of 0.2. $\mathcal{N}_i$ is the number of adjacent nodes of i.

Taking interaction of the left hand as an example, we calculate the interaction between the left hand and the major body, right hand, and object:

$$\hat{\boldsymbol{X}}_l = \sigma(\alpha_{r,l}\boldsymbol{W}_{r,l}\tilde{\boldsymbol{X}}_r^{''} + \alpha_{b,l}\boldsymbol{W}_{b,l}\tilde{\boldsymbol{X}}_b^{''} + \alpha_{l,s}\boldsymbol{W}_{l,s}\tilde{\boldsymbol{S}}), \tag{9}$$

where $\boldsymbol{W}_{r,l}$, $\boldsymbol{W}_{b,l}$, $\boldsymbol{W}_{l,s}$ are learnable weight. Also, we can obtain the other human pose features $\hat{\boldsymbol{X}}_b$ and $\hat{\boldsymbol{X}}_r$. The resulting features $\{\hat{\boldsymbol{X}}_l, \hat{\boldsymbol{X}}_b, \hat{\boldsymbol{X}}_r\}$ are then followed by a predictor composed of a MLP and IDCT to regress the final features into the predicted sequence $\{\hat{\boldsymbol{Y}}_{l,b,r}\}$.

3.5. Training Losses

Joint Loss. We first use L_2 loss to calculate the average error of each predicted joint [41, 50, 62]:

$$\mathcal{L}^p = \frac{1}{N\Delta T}\sum_{n=1}^{N}\sum_{t=1}^{\Delta T}||\hat{\boldsymbol{p}}_{n,t} - \boldsymbol{p}_{n,t}||_2, \tag{10}$$

where $\hat{\boldsymbol{p}}_{n,t}$ denotes the predicted n-th joint in t-th frame, $\boldsymbol{p}_{n,t}$ is the corresponding ground truth. $N = N_b + N_l + N_r = 55$ is the number of joints of the whole body.

Bone Length Loss. To further account for the connectivity within the human body, we introduce the bone length loss as a physical constraint [14, 33, 44]:

$$\mathcal{L}^b = \frac{1}{(N-1)\Delta T}\sum_{n=1}^{N-1}\sum_{t=1}^{\Delta T}\|\hat{\boldsymbol{b}}_{n,t} - \boldsymbol{b}_{n,t}\|_1, \tag{11}$$

where $\hat{\boldsymbol{b}}_{n,t}$ denotes the length of the n-th bone in t predicted frame, $\boldsymbol{b}_{n,t}$ is the corresponding ground truth.

Distance Loss. As the distance changes, the interaction strength of hand-object modal features also varies. To obtain more precise predictions, we use a distance loss ensure the reasonable distance between the hands and the object:

$$\mathcal{L}_l^d = \frac{1}{\Delta T}\sum_{t=1}^{\Delta T}\left\|\hat{\boldsymbol{d}}_{l\leftrightarrow o,t} - \boldsymbol{d}_{l\leftrightarrow o,t}\right\|_1, \tag{12}$$

where $\hat{\boldsymbol{d}}_{l\leftrightarrow o,t}$ refers to the minimum distance between the 15 joints of the fingers to the object in t frame. Similarly, we can also calculate $\mathcal{L}_b^d$ and $\mathcal{L}_r^d$, to form the final distance loss $\mathcal{L}^d = \frac{1}{3}(\mathcal{L}_l^d + \mathcal{L}_b^d + \mathcal{L}_r^d)$.

	Body Parts	Major body				Left Hand				Right Hand				Left Hand (AW)				Right Hand (AW)				Whole Body			
	Time(sec)	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0
w/o object	LTD [44]	8.7	**18.9**	39.2	48.7	19.7	57.0	143.3	181.5	33.3	77.5	159.1	195.6	9.1	18.3	33.8	41.4	17.2	28.3	46.1	53.1	18.3	45.6	101.5	126.1
	DMGNN [36]	11.2	23.1	43.8	53.5	24.8	62.0	153.2	190.1	38.1	83.0	166.7	205.7	10.0	21.7	39.1	44.4	21.6	32.6	49.7	60.5	23.0	55.7	107.7	131.4
	PGBIG [40]	10.4	21.7	44.0	52.8	22.8	61.5	149.9	186.7	37.6	82.4	164.5	203.9	10.5	22.2	38.7	43.5	21.5	31.1	48.8	58.7	22.6	53.6	104.3	129.6
	SPGSN [35]	9.3	21.0	43.2	52.6	25.3	61.1	129.8	164.2	37.2	81.5	165.9	202.8	9.3	18.5	34.0	41.6	16.1	28.8	49.3	56.9	21.2	48.4	100.3	124.0
w/ object	LTD [44]	**8.5**	20.3	38.5	47.3	**18.9**	56.9	140.5	177.4	32.6	76.4	157.5	187.0	**9.0**	17.8	35.6	43.7	17.1	27.8	47.8	52.0	**17.9**	44.8	98.4	119.7
	DMGNN [36]	10.0	22.2	42.7	50.5	23.3	60.9	148.9	188.4	36.7	81.7	163.0	187.0	9.7	20.4	37.4	42.2	19.4	31.8	47.3	57.5	21.1	52.7	102.6	126.4
	PGBIG [40]	10.1	19.9	41.6	49.5	20.6	57.8	144.3	178.5	37.4	80.3	157.7	194.2	9.8	21.6	36.6	41.7	20.8	30.6	47.3	55.8	21.1	50.2	98.8	121.4
	SPGSN [35]	9.0	20.3	41.5	50.5	23.6	58.9	126.5	157.5	35.4	79.3	140.4	169.3	**9.0**	**17.1**	**32.6**	**38.7**	15.7	26.6	46.8	53.4	20.0	46.1	95.7	120.0
	C^3HOST	8.7	**18.9**	**37.5**	**46.1**	26.5	**56.5**	**117.9**	**153.3**	**29.6**	**66.0**	**128.5**	**161.5**	12.4	21.0	35.5	42.7	**14.9**	**26.4**	**42.1**	**49.4**	19.2	**42.0**	**84.3**	**106.8**

Table 1. Comparison of the average results of all actions on the GRAB dataset [53]. The best results are highlighted in bold.

MMD Loss. To alleviate the features heterogeneity of body's internal context, we utilize the MMD loss in Eq. 3 as an alignment constraint [9, 18].

$$\mathcal{L}^{MMD} = \text{MMD}(\boldsymbol{X}'_l, \boldsymbol{X}'_b) + \text{MMD}(\boldsymbol{X}'_b, \boldsymbol{X}'_r) + \text{MMD}(\boldsymbol{X}'_r, \boldsymbol{X}'_l). \quad (13)$$

Final Loss, is the weighted sum of the above losses:

$$\mathcal{L} = \lambda_1 \mathcal{L}^p + \lambda_2 \mathcal{L}^b + \lambda_3 \mathcal{L}^d + \lambda_4 \mathcal{L}^{MMD}, \quad (14)$$

where λ_1, λ_2, λ_3, λ_4 are the trade-off parameters.

3.6. Implementation Details

We utilize 4 NVIDIA RTX 3090 GPUs with the Distributed Data Parallel (DDP) training approach and train the model using the AdamW optimizer. The weight decay is set to the default value of 0.01. A total of 50 epochs are conducted, with a batch size of 64 on each GPU, an initial learning rate of 0.001×4 on each GPU, and a learning rate decay of 0.96 every two epochs. To prevent the occurrence of overfitting during training, every layer of the entire model was equipped with a Batch Normalization (BN) layer, and a dropout of 0.5 was applied. The trade-off parameters $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ are set as $\{1, 0.1, 0.1, 0.1\}$

4. Experiments

4.1. Datasets

Dataset-1: GRAB [53] is a recently released dataset with $\approx 1.6M$ frames of 10 actors performing a total of 29 actions. It annotates the whole-body SMPL-X parameters [47] using high-precision MoCap techniques, in which 25 joints (3D coordinates) are defined as the body ($N_b = 25$), and each hand is denoted as 15-joints ($N_l = N_r = 15$). Consistent with [16, 17, 32], we down-sample all sequences to 30 fps, and remove both start and end T-pose (1 second). The length of the observation and prediction is equal, *i.e.*, $T = \Delta T = 30$ (1 second). We split subject-10 (S10) as the test, S2-S9 as the training, and S1 as the validation set. The GRAB dataset includes 50 objects. We extract its point cloud data and downsample the number of vertices of the object mesh to $O = 1024$ as the shape feature.

Time(sec)	0.2	0.4	0.8	1.0	Average
LTD [44]	**21.0**	48.7	106.9	131.7	77.1
PGBIG [40]	24.2	55.3	112.6	140.1	83.0
SPGSN [35]	21.8	48.1	108.1	135.6	78.4
C^3HOST	21.4	**45.9**	**97.5**	**114.1**	**69.7**

Table 2. Comparison of the average results of whole body on the BEHAVE dataset (*w/* object).

Dataset-2: BEHAVE dataset [8] includes 386 samples captured over 15,000 frames using 4 Kinect RGB-D cameras at a frame rate of 30 fps. The dataset involves the observation of 17 types of Human-Object Interactions (HOI) across 8 subjects and 20 objects. Each pose is represented by a 67-joint skeleton, with 25 joints for the body and 21 for each hand. According to the official documentation, the dataset is divided into training subsets consisting of 231 sequences and testing subsets comprising 90 sequences. However, it is noteworthy that 295 out of the total 321 sequences contain fewer than 60 frames.

4.2. Baselines and Evaluation Metrics

Baselines: Due to the absence of whole-body motion prediction methods, we compare our approach with the standard ones, including LTD [44], DMGNN [36], PGBIG [40], SPGSN [35]. LTD [44] proposes a GCN with a learnable adjacency matrix. DMGNN [36] upsamples human features and constructs a GCN-based multi-scale feature pyramid. PGBIG [40] designs a two-stage prediction framework, where the initial network provides a 'initial guess' for the main network. SPGSN [35] introduces an adaptive graph scattering technique. For an unbiased comparison, we use the following strategies: (1) We expand the predicted joint count from 17 or 25 to 55, encompassing the body, left hand, and right hand. This demonstrates the effectiveness of cross-context alignment. (2) We introduce object information into the network input while disregarding heterogeneous external cross-modal information. This affirms the effectiveness of human-object interaction learning. The baselines are retrained according to the aforementioned strategies, while maintaining consistency with other aspects. It is confirmed that ours re-trained baseline out-

	Action	A1 cook				A2 eat				A3 drink				A4 lift				A5 wear				A6 squeeze			
	Time(sec)	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0
Major Body	LTD [44]	13.6	**28.9**	48.1	55.4	12.1	28.0	56.2	71.7	12.2	23.4	**37.6**	**40.1**	7.9	20.3	45.4	54.9	**6.3**	14.6	29.3	35.3	5.6	12.7	22.2	**26.6**
	DMGNN [36]	16.4	30.2	49.5	60.2	17.9	37.6	68.9	88.4	14.5	32.1	44.8	58.0	12.1	26.3	51.1	61.5	13.2	38.0	54.8	65.5	23.1	48.1	64.2	75.4
	PGBIG [40]	14.9	30.4	48.7	54.8	17.5	36.5	66.4	83.2	15.7	30.2	42.6	53.2	11.4	24.3	48.9	62.4	12.1	28.7	46.9	57.7	21.5	46.2	60.3	72.4
	SPGSN [35]	**13.2**	29.1	46.9	53.4	18.4	34.8	65.1	82.0	15.6	28.9	40.0	48.6	10.6	22.6	44.0	51.6	7.4	16.9	28.8	36.1	7.9	13.8	23.6	27.9
	C^3HOST	14.4	29.3	**46.8**	**52.7**	**7.2**	**15.2**	**27.4**	**32.4**	**6.9**	**16.6**	37.9	48.1	**7.3**	**15.5**	**30.5**	**37.6**	6.5	**14.4**	**27.9**	**34.0**	**5.4**	**11.2**	**21.5**	27.4
Left hand	LTD [44]	**22.3**	**60.3**	131.6	164.8	**21.7**	**52.5**	129.4	187.5	51.8	123.4	189.9	185.8	21.0	66.3	145.7	163.8	**22.3**	67.7	155.3	187.8	**18.1**	35.6	**41.9**	**54.6**
	DMGNN [36]	27.7	64.9	135.8	174.6	38.6	87.5	178.3	234.4	56.2	128.8	210.5	265.4	22.2	68.7	152.5	181.2	34.1	70.6	175.3	200.2	24.1	49.3	60.2	73.1
	PGBIG [40]	26.5	66.3	133.7	169.5	36.9	88.2	156.2	225.6	56.7	126.4	215.7	264.6	23.1	66.9	154.0	178.4	33.0	69.2	169.3	193.8	23.0	45.6	56.7	72.4
	SPGSN [35]	24.8	61.1	129.7	150.5	36.5	94.6	206.1	263.6	51.4	119.8	219.3	242.7	**20.3**	65.2	152.9	175.5	27.7	68.6	160.4	172.5	22.9	47.3	55.2	70.3
	C^3HOST	26.8	**60.3**	**117.7**	**143.6**	28.5	62.6	**113.5**	**130.9**	**18.3**	**46.4**	**114.2**	**153.8**	28.1	**56.2**	**111.1**	**142.0**	30.2	**65.1**	**122.3**	**148.4**	18.3	**30.8**	53.3	71.6
Right hand	LTD [44]	**55.8**	126.0	185.6	215.5	35.3	79.3	152.1	204.3	22.9	82.2	171.7	**167.2**	25.5	81.5	192.9	229.1	**17.7**	**53.8**	133.3	152.5	25.1	47.8	**81.0**	**93.7**
	DMGNN [36]	62.2	136.5	208.6	239.5	37.5	78.3	156.5	215.0	23.5	85.8	211.6	221.4	27.3	83.4	194.5	231.0	33.3	89.6	177.6	210.2	26.4	54.2	92.1	103.7
	PGBIG [40]	59.7	131.8	199.3	220.2	34.2	76.4	152.3	212.5	24.0	87.6	208.3	210.5	26.1	82.5	195.7	233.7	30.6	81.3	173.2	199.9	25.7	54.0	91.4	102.5
	SPGSN [35]	60.2	120.5	146.7	156.7	31.8	59.5	146.3	207.6	22.5	92.0	213.3	249.4	**21.3**	76.4	185.5	215.6	25.6	65.4	140.8	182.1	24.7	52.6	89.7	98.7
	C^3HOST	60.1	**118.1**	**145.0**	**155.0**	**22.0**	**51.6**	**107.3**	**140.7**	**18.7**	**51.1**	**124.8**	171.1	26.1	**54.1**	**99.9**	**124.0**	24.0	61.1	**121.0**	**151.6**	**21.9**	**45.6**	81.5	101.6
Whole body	LTD [44]	**28.7**	66.6	112.7	133.9	21.0	48.7	102.3	139.4	25.9	66.7	115.7	114.5	16.3	49.5	113.0	132.1	**11.8**	**31.1**	87.9	112.3	14.3	28.5	**43.6**	**52.6**
	DMGNN [36]	33.5	71.4	120.9	145.8	29.4	60.2	129.0	170.6	30.0	74.4	138.3	163.7	18.2	54.3	118.9	144.3	22.5	51.0	79.8	91.0	18.4	36.6	58.3	62.3
	PGBIG [40]	31.6	70.6	117.4	136.2	25.3	58.3	118.9	169.3	28.2	72.1	131.5	159.9	17.1	51.6	116.5	136.4	21.4	49.4	76.3	85.6	17.4	35.7	54.2	61.0
	SPGSN [35]	30.4	65.4	101.0	112.9	27.0	57.9	125.7	165.8	27.2	70.9	136.2	156.3	**16.2**	48.9	112.3	130.1	13.7	36.8	89.5	112.2	16.6	33.5	50.2	58.8
	C^3HOST	30.3	**62.0**	**86.1**	**105.3**	**17.0**	**38.0**	**72.6**	**88.8**	**13.2**	**34.2**	**82.4**	**110.5**	18.1	**37.1**	**71.4**	**89.6**	17.7	41.0	**79.0**	**97.3**	**13.4**	**26.0**	46.5	60.0

Table 3. Comparison of the detailed results of each action categories on the GRAB dataset.

performs the publicly available baseline one (7.6mm error reduction), indicating the fairness of the re-training strategy.

Evaluation Metrics: In accordance with previous work [14, 17, 36, 44], we use mean per joint position error (MPJPE) [28] measured in mm to assess the 3D prediction accuracy. However, MPJPE is unable to capture subtle movements and semantic information in the hands. Therefore, we utilize wrist joint alignment with hand joints, and calculate the MPJPE-AW [17].

4.3. Results Analysis

C^3HOST v.s. Baselines. Table 1 shows a quantitative comparison between the baseline models and our approach. In scenarios without interacting with objects, our results outperform the baseline model in most cases. Two distinct data trends can be observed: (1) The superiority of our method becomes more pronounced with an increase in prediction time. This is because, in short-term predictions, other methods can rely on the inertia of movements and spatial continuity of joints to achieve good predictive results. However, over time, this inertia and spatial continuity become less evident and may even produce adverse effects. At this point, as the human body's movements become more closely intertwined with external interactions with objects, our approach can extract partial action intentions from these external interactions, thereby achieving accurate long-term predictions. (2) Predictions of right-hand movements are more accurate compared to those of left-hand and body movements. The prevalence of right-hand interactions with objects in the GRAB is the contributing factor to this observation. It also indicates that our approach can effectively extract interaction information between humans and objects. When interacting with objects, it is evident that all baselines showed improved performance,and our approach still outperforms the comparative methods. This is attributed to the fact that other comparative methods neglect the heterogeneity of human internal-context and human-object modalities, making them ineffective in learning interaction information. This indicates that our proposed cross-context alignment and human-object interaction learning demonstrate a significant impact in aligning heterogeneous features, and facilitates cross-context and cross-modal interaction learning. We present the results of validation experiments involving object interactions on the BEHAVE dataset in Table 2. In the majority of cases, our method consistently produces the smallest errors. These findings affirm that our proposed approach demonstrates strong generalization and robustness across diverse datasets.

Comparison of Specific Actions. To validate the compatibility of our method, we select the 6 common daily actions. Table 3 reports more detailed prediction results for these six action sets. We observe that our method outperforms the baseline approaches in most cases, which evidences the effectiveness of enhancing human motion prediction through the extraction of human-object interaction features. Furthermore, the compatibility of our method has been validated by considering various motion patterns and interacting objects.

Visualization. As shown in Figure 3, we present qualitative results of the 'teapot-pour' and 'banana-eat' actions. We select the LTD, which performs the best in the baseline models, for comparison. However, our method shows closer predictions to the ground truth for the upper limbs and hands between 0.8s and 1.0s, as evident in the detailed

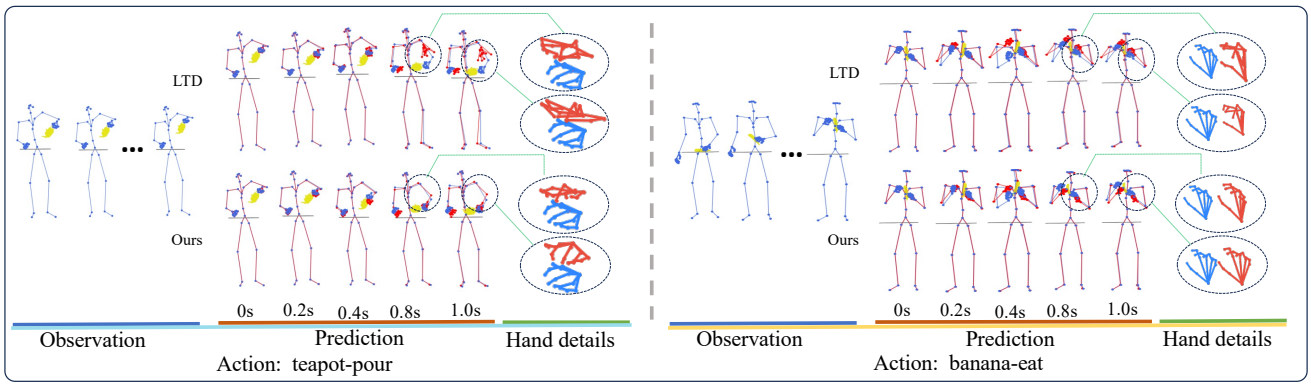

Figure 3. Visualizations of the predicted whole-body human poses. In each sub-figure, the left is past observed sequence, the middle are predicted poses, and the right, within the circles, are magnified hand details. The prediction and ground truth are represented by red and blue skeletons, respectively. Objects are depicted as yellow mesh images. We observe from the hand detail that compared to the ground truth, the baseline method distorts the hand joints, whereas our method accurately restore the hand movements.

CCA	CCI	0.2s	0.4s	0.8s	1.0s	Average
		19.2	44.5	92.0	117.0	68.2
	✓	**19.0**	42.6	88.8	111.3	65.4
✓	✓	19.1	**42.0**	**87.3**	**110.7**	**64.8**

Table 4. Ablation experiment of internal interaction.

CMA	HOIL	gated sharing unit	0.2s	0.4s	0.8s	1.0s	Average
			19.1	42.0	87.3	110.7	64.8
	✓		**17.1**	**41.6**	89.8	113.9	65.6
✓	✓		18.3	42.5	86.6	111.1	64.6
✓	✓	✓	19.2	42.0	**84.3**	**106.8**	**63.0**

Table 5. Ablation experiment of external interaction.

fine-grained hand prediction images. This indicates that incorporating information from objects interacted with by human can effectively enhance the results of whole-body (especially hands) movements. It also validates the importance of the collaborative analysis of cross-motion internal context and cross-modal external interaction for the forecasting of 3D whole-body human poses with grasping objects.

4.4. Ablation Studies

We conduct ablation experiments on the model to study the impact of several key components on C^{3}HOST. The facilitation of various components in internal interactions are confirmed in Table 4. The average error is 68.2mm when diverse motion contexts are not considered. When considering only context interaction, the average error decreases to 65.4mm, indicating a noticeable improvement in performance. Simultaneously considering the alignment and interaction of context, the average error further decreases to 64.8mm. This suggests that context interaction can enhance model performance, with interaction being relatively more important than alignment.

The result in Table 5 indicates the impact of objects on the model. When object information is not considered, the average error is 64.8mm. Directly incorporating object shape information results in a noticeable increase in the average error to 65.6mm. This indicates significant differences in the feature distribution between objects and human motion joints, and direct feature fusion can lead to feature contamination. After aligning object features with human joints features, the average error decreases to 64.6mm. Moreover, with the addition of gated sharing unit, the average error further decreases to 63.0mm.

5. Conclusion

In this work, we introduce a novel framework to address the challenge of forecasting 3D whole-body human poses with grasping objects. It is designed to refine the alignment and integration of features across different contexts and modalities. We incorporates gated human-object distance information, which establishes the interaction strength between human and object features over time. This temporal interaction modeling is essential for accurately capturing the dynamic nature of human-object interactions. Compared to conventional models, our method fosters enhanced cross-contextual intra-body and cross-modal human-object feature interactions. It emphasizes the importance of considering the body's internal dynamics and the external environment's influence on human motion. Our comprehensive set of experiments confirms the efficacy of this novel approach. This advancement marks a major step forward in the field of human motion prediction, particularly in scenarios where detailed interaction with objects is critical.

Acknowledgements

This work was supported in part by the Natural Science Foundation of Jiangsu Province (BK20220939) and the National Natural Science Foundation of China (62306141).

References

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

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

[3] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974. 3

[4] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE, 2021. 1

[5] Gopalakrishnan Anand, Mali Ankur, Kifer Dan, Lee Giles C., and Ororbia Alexander. A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[6] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1418–1427, 2018. 1

[7] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: Probabilistic 3D Human Motion Prediction via GAN. In *CVPR*, pages 1418–1427, 2018. 2

[8] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15935–15946, 2022. 6

[9] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 2, 6

[10] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2949, 2016. 2

[11] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019. 2

[12] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6992–7001, 2020. 2, 3

[13] Qiongjie Cui and Huaijiang Sun. Towards accurate 3d human motion prediction from incomplete observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4801–4810, 2021. 1, 2

[14] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6519–6527, 2020. 2, 3, 5, 7

[15] 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, 2021. 2

[16] Christian Diller, Thomas Funkhouser, and Angela Dai. Forecasting characteristic 3d poses of human actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15914–15923, 2022. 6

[17] Pengxiang Ding, Qiongjie Cui, Min Zhang, Mengyuan Liu, Haofan Wang, and Donglin Wang. Expressive forecasting of 3d whole-body human motions. *arXiv preprint arXiv:2312.11972*, 2023. 6, 7

[18] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015. 2, 6

[19] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4346–4354, 2015. 2

[20] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, pages 1–12, 2023. 2

[21] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and José MF Moura. Few-shot human motion prediction via meta-learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 432–450, 2018. 1

[22] Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, José MF Moura, and Manuela Veloso. Teaching robots to predict human motion. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 562–567. IEEE, 2018. 1

[23] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13053–13064, 2022. 3

[24] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4809–4819, 2023. 1, 4

[25] Ankur Gupta, Julieta Martinez, James J Little, and Robert J Woodham. 3d pose from motion for cross-view action recognition via non-linear circulant temporal encoding. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2601–2608, 2014. 1

[26] Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. 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)*, pages 221–231, 2017. 4

[27] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019. 2

[28] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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. 7

[29] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5308–5317, 2016. 2

[30] Nath Kundu Jogendra, Gor Maharshi, and Venkatesh Babu R. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *AAAI*, 2019. 1

[31] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[32] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11585–11594, 2021. 6

[33] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5226–5234, 2018. 1, 2, 4, 5

[34] Maosen Li, Siheng Chen, Zihui Liu, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton graph scattering networks for 3d skeleton-based human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 854–864, 2021. 1, 4

[35] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3d human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 18–36. Springer, 2022. 1, 2, 6, 7

[36] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. 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 (CVPR)*, pages 214–223, 2020. 2, 3, 6, 7

[37] Paul Pu Liang, Peter Wu, Liu Ziyin, Louis-Philippe Morency, and Ruslan Salakhutdinov. Cross-modal generalization: Learning in low resource modalities via meta-alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2680–2689, 2021. 2

[38] Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8148–8156, 2021. 2

[39] Zhenguang Liu, Kedi Lyu, Shuang Wu, Haipeng Chen, Yanbin Hao, and Shouling Ji. Aggregated multi-gans for controlled 3d human motion prediction. *AAAI*, 2021. 1

[40] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6437–6446, 2022. 1, 6, 7

[41] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6437–6446, 2022. 2, 5

[42] Odalric-Ambrym Maillard, Timothy A Mann, and Shie Mannor. How hard is my mdp?" the distribution-norm to the rescue". *Advances in Neural Information Processing Systems*, 27, 2014. 2

[43] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Proceedings of the European conference on computer vision (ECCV)*, pages 474–489. Springer, 2020. 1

[44] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9489–9497, 2019. 2, 3, 4, 5, 6, 7

[45] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. Autodial: Automatic domain alignment layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5067–5075, 2017. 5

[46] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2891–2900, 2017. 1, 2

[47] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 6

[48] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3, 4

[49] Adrià Recasens, Jason Lin, João Carreira, Drew Jaegle, Luyu Wang, Jean-baptiste Alayrac, Pauline Luc, Antoine Miech,

Lucas Smaira, Ross Hemsley, et al. Zorro: the masked multimodal transformer. *arXiv preprint arXiv:2301.09595*, 2023. 2

[50] Alejandro Hernandez Ruiz, Juergen Gall, and Francesc Moreno-Noguer. Human Motion Prediction via Spatio-Temporal Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7134–7143, 2018. 1, 2, 5

[51] Alexander J Smola, A Gretton, and K Borgwardt. Maximum mean discrepancy. In *13th international conference, ICONIP*, pages 3–6, 2006. 4

[52] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13263–13273, 2022. 2

[53] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Proceedings of the European conference on computer vision (ECCV)*, pages 581–600. Springer, 2020. 6

[54] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 2

[55] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 2, 5

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

[57] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In *The world wide web conference*, pages 2022–2032, 2019. 2

[58] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. *arXiv preprint arXiv:1905.07718*, 2019. 2

[59] Mathieu Salzmann Wei Mao, Miaomiao Liu. Generating Smooth Pose Sequences for Diverse Human Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 474–489, 2021. 2

[60] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *Proceedings of the European conference on computer vision (ECCV)*, pages 257–274. Springer, 2022. 2

[61] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 251–269. Springer, 2022. 2

[62] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Proceedings of the European conference on computer vision (ECCV)*, 2020. 1, 5

[63] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018. 2

[64] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *Proceedings of the European conference on computer vision (ECCV)*, pages 676–694. Springer, 2022. 1, 2, 3

[65] Liu Zhenguang, Wu Shuang, Jin Shuyuan, Tang Minghua, Lu Shijian, Zimmermann Richard, and Cheng Lichen. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1