Forecasting Bimanual Object Manipulation Sequences from Unimanual Observations

Haziq Razali, Yiannis Demiris*

Personal Robotics Lab, Dept. of Electrical and Electronic Engineering,
Imperial College London
{h.bin-razali20,y.demiris}@imperial.ac.uk

Abstract

Learning to forecast bimanual object manipulation sequences from unimanual observations has broad applications in assistive robots and augmented reality. This challenging task requires us to first infer motion from the missing arm and the object it would have been manipulating were the person bimanual, then forecast the human and object motion while maintaining hand-object contact during manipulation. Previous attempts model the hand-object interactions only implicitly, and thus tend to produce unrealistic motion where the objects float in air. We address this with a novel neural network that (i) identifies and forecasts the pose for only the objects undergoing motion through an object motion module and (ii) refines human pose predictions by encouraging hand-object contact during manipulation through an ensemble of human pose predictors. The components are also designed to be generic enough for use in both unimanual and bimanual contexts. Our approach outperforms the state-of-the-art pose forecasting methods on bimanual manipulation datasets.

Introduction

Forecasting bimanual object manipulation has broad applications in extended reality (García et al. 2022) and human-robot interaction (Koppula and Saxena 2013). While forecasting human motion has been heavily studied in the literature, very few include the object motion in its prediction (Corona et al. 2020; Razali and Demiris 2023). Being able to forecast the movement of both hands as one reaches a fork and another a spoon is useful as it opens up additional use cases such as occlusion handling or vision-based rehabilitative robots (Lum et al. 2002). For instance, a fixed UR10 robot can help perform daily tasks in the kitchen by acting as the second arm for a patient who is suffering from single arm paralysis or a temporary injury to their upper limb.

We thus investigate in this paper the forecasting of bimanual object manipulation sequences from unimanual observations, where we receive motion data that has been appropriately processed to make it appear as if the person is one-handed, and reconstruct the most likely configuration before forecasting (Figure 1). Previous research has focused on developing more sophisticated Graph Recurrent Neural Networks (GRNNs) to tackle reconstruction (Cui and Sun 2021) or forecasting (Corona et al. 2020; Razali and Demiris 2023). While these efforts have led to lower errors, they do not account for the semantics of bimanual object manipulation, in that the hand remains in contact with the object during manipulation and only objects undergoing either direct or indirect manipulation can have non-zero velocities. One major reason why these semantics are not captured is that the human or object pose is predicted directly after the output of the GRNN with an MLP. As such, although the features are extracted jointly through a graph, the prediction of each node is ultimately weakly or implicitly conditioned on all others. The result of this is unrealistic hand-object motion where the objects drift and float in air. This also leads to an issue where objects that are supposed to be stationary are inadvertently predicted a non-zero velocity due to the MLP's inability to output exact zeros for the object pose velocities, resulting in objects that drift.

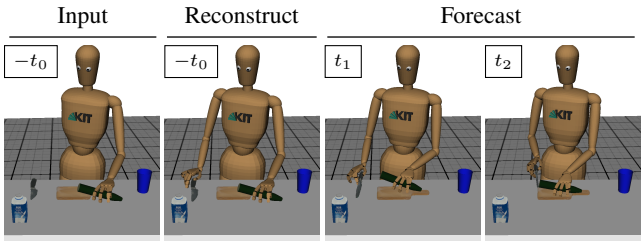

Figure 1: Given a unimanual input, we reconstruct and forecast the sequence as if it were performed bimanually.

In this paper, we demonstrate the importance of explicitly incorporating these semantics in the prediction process through the incorporation of two key components. First, we introduce an object motion module that outputs the motion and binary probability of each object undergoing motion. These probabilities enable the model to effectively zero-out velocities for objects not undergoing manipulation, eliminating object drift. Second, we introduce a human pose ensemble module that refines the predicted human pose, explicitly conditioned on the object each hand is in contact with. This refinement encourages hand-object contact during manipu-

*Yiannis Demiris is supported by a Royal Academy of Engineering Chair in Emerging Technologies.
Code at www.imperial.ac.uk/personal-robotics/software

lation, making the predictions more plausible. We couple our contributions with a generic GRNN for feature processing to show through experiments that the complexity of such networks (Corona et al. 2020; Cui and Sun 2021; Razali and Demiris 2023) do not necessarily result in better or more plausible outputs. Lastly, our components are also designed as stand alone units generic enough for use in any existing networks, and in both unimanual and bimanual contexts.

Overall, our contributions are as follows: (1) We tackle the novel task of forecasting bimanual object manipulation sequences by observing the interaction to be unimanual. (2) We propose an object motion module to zero-out unnatural object drift. (3) We propose a human pose ensemble module that encourages hand-object contact during manipulation. (4) Although we use off-the-shelf components, our novel way of combining them lets us achieve state-of-the-art performance on bimanual manipulation datasets.

Related Work

Human Pose Forecasting: Typical methods treat human pose forecasting as a sequence to sequence learning problem, differing mainly in their encoding-decoding strategies, employing either autoregressive methods (Martinez, Black, and Romero 2017; Corona et al. 2020) or non-autoregressive ones (Cui and Sun 2021). An early work (Martinez, Black, and Romero 2017) showed the benefit of predicting pose velocities instead of positions in order to eliminate the discontinuity between the first predicted frame and the last observed frame which was heavily adopted by successive works. These works can also be categorized into deterministic (Martinez, Black, and Romero 2017) or stochastic (Liu et al. 2021), using Variational Autoencoders (VAEs) (Kingma and Welling 2013) or Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) respectively, with the design choice hinging on whether there is sufficient variation to be learnt by the model. Many recent works incorporate additional context such as scene (Corona et al. 2020), eye gaze (Razali and Demiris 2022b; Zheng et al. 2022), or object coordinates (Razali and Demiris 2022a).

Most similar to our work is the context-aware model by (Corona et al. 2020) and the multi-task Graph Convolutional Network by (Cui and Sun 2021). (Corona et al. 2020) leverages Graph Networks to incorporate both spatial and temporal context for human and object motion prediction. (Cui and Sun 2021) addresses the challenge of predicting human motion from incomplete observations. However, there are notable distinctions. (Corona et al. 2020) primarily focused on generic human-object interaction using fully observed data. Furthermore, the architecture does not encourage contact between the hands and objects, resulting in unrealistic motion where the objects float in space. We show in this paper that incorporating a separate neural network tasked specifically to encourage contact results in an improvement. (Cui and Sun 2021) forecasts the human pose from incomplete observations except they mask the pose in an unstructured manner and deal only with the human pose. Ours can be thought of as complementary wherein the observation is incomplete in the sense that it is performed unimanually.

Hand and Human Pose Synthesis: Methods developed for synthesizing human motion from scratch can be categorized on the conditioning variable: scene (Hassan et al. 2021), audio (Li et al. 2021), text (Tevet et al. 2023; Jiang et al. 2023), object (Taheri et al. 2020, 2022; Razali and Demiris 2023), or an action label (Starke et al. 2019; Petrovich, Black, and Varol 2021; Razali and Demiris 2023). Hassan et al. (Hassan et al. 2021) learns the contact probability of each vertex of the human mesh to objects such as floor and sofa to enforce contact during generation. (Li et al. 2021) uses a cross-modal Transformer to learn the correlation between motion and music. Works that generate a single frame of the full body or the human hand given the object mesh (Taheri et al. 2020) often assume there to be little to no in-hand manipulation. Subsequent works synthesize a human approaching, grasping, and manipulating a lone object (Taheri et al. 2022). Our problem in contrast, requires us to manipulate different objects simultaneously that interact with each other in the presence of distractors.

In our previous work (Razali and Demiris 2023), we synthesized the human and object motion from start to finish conditioned on an action label, through a modularized architecture that leverages the varying degree each joint is involved during object manipulation. In particular, we modeled the hand and object motion using a GRNN, and utilized separate RNNs to reconstruct the body and finger joints based on the generated wrist and object motion. In contrast, our current task involves forecasting without the aid of an action label, is not confined to the start frame, while dealing with incomplete observations. Furthermore, (Razali and Demiris 2023) suffers from the above-mentioned problem of objects that float in space. Lastly, a similarity amongst (Corona et al. 2020; Cui and Sun 2021; Razali and Demiris 2023) is the use of an MLP as the final layer to predict either the human or object pose. This leads to an issue where objects not undergoing either direct or indirect manipulation are inadvertently predicted a non-zero velocity, resulting in objects that drift. We address this issue in our work.

Method

We let $\bar{y}^{-T_1:0} = [\bar{y}^{-T_1}, ..., \bar{y}^0] \in R^{T_1 \times J \times 3}$ denote the input human pose with J joints across T_1 timesteps, $\bar{x}^{-T_1:0} = \{\bar{x}_1^{-T_1:0}, ..., \bar{x}_N^{-T_1:0}\} \in R^{T_1 \times N \times K \times 3}$ is the set of N objects present in the scene represented by K motion capture markers or its centroid, $l = [l_1, ..., l_N, l_{\text{human}}] \in R^{(N+1)\times(N+1)}$ are the one-hot labels, and where the data has been processed by a masking function to make it appear as if the interaction is unimanual. We mask the data by selecting one arm and the objects the selected arm interacts with at every timestep. We then set the coordinates throughout time of the entire arm to the zero vector and the object to its initial pose. That way, when the sequence is visualized, it would appear as if the person is moving only the other arm to interact with one object at a time. This operation can be expressed as $\bar{y}^{-T_1:0}, \bar{x}^{-T_1:0} = \text{mask}(y^{-T_1:0}, x^{-T_1:0})$.

Given the unimanual sequence, our objective is to fore-

cast the motion as though it were performed by an individual who is bimanually capable $p(y^{0:T_2}, x^{0:T_2}|\bar{y}^{-T_1:0}, \bar{x}^{-T_1:0})$. We factorize this task into two sequential stages. We first reconstruct the most likely bimanual sequence $p(y^{-T_1:0}, x^{-T_1:0}|\bar{y}^{-T_1:0}, \bar{x}^{-T_1:0})$ then forecast its motion $p(y^{0:T_2}, x^{0:T_2}|y^{-T_1:0}, x^{-T_1:0})$. Figure 2 illustrates the framework of our system. In the following, we describe it in more detail.

Bimanual Reconstruction

During bimanual interaction, the human hands are often strongly correlated to each other and the objects' motion. For instance, the left hand of a right-handed person may hold a cup in place while the right pours water from a bottle. We leverage this by using a generic Graph Gated Recurrent Unit (GRU) as the encoding backbone to extract a shared representation of each entity at every timestep. Specifically, we initialize a densely connected graph where each vertex $v_j^t = \varphi([\bar{x}_i^t, l_i])$ or $\varphi([\bar{y}^t, l_{\text{human}}])$, is an encoding of the pose and label via the MLP φ, and the edge the numerical subtraction between two neighbouring vertices. Henceforth, we use $j \in [1, ..., N, h]$ when indexing the objects and human, and i the object only. We then obtain the shared representation for $t = [-T_1, ..., 0]$ as follows:

$$g_j^t = F_{\text{enc}} = \text{GRU}(\max_{j' \in \mathcal{N}(j)} \varphi([v_j^t, v_{j'}^t - v_j^t]); g_j^{t-1}) \quad (1)$$

Next, studies have shown that humans subconsciously tend to visualize and plan out the object motion before motor movement (Johansson and Cole 1992). In the same spirit, we further factorize the reconstruction into two sub-tasks that we explain in the following subsections: object prediction $p(x_{-T_1:0}|\bar{y}_{-T_1:0}, \bar{x}_{-T_1:0})$ followed by human pose reconstruction $p(y_{-T_1:0}|x_{-T_1:0}, \bar{y}_{-T_1:0})$.

Object Motion Module (OMM): We can often make the assumption during bimanual object manipulation that each hand can manipulate a maximum of one object. We can also categorize an object motion into two states: stationary or potentially moving either when undergoing manipulation, either directly or indirectly, such as an egg rolling after being released from a hand. This gives rise to our multitask OMM that predicts the object pose, weighted by their binary probability scores:

$$\alpha_i^t = \sigma(\varphi_{c_1}([g_i^t, l_i])) \quad (2)$$

$$\hat{x}_{\varphi,i}^t = \varphi_p([g_i^t, l_i]) \quad (3)$$

$$\hat{x}_i^t = \hat{\alpha}_i^t \hat{x}_{\varphi,i}^t + (1 - \hat{\alpha}_i^t) x_i^{t-1} \quad (4)$$

where σ is the sigmoid function and α_i^t is a scalar probability. Eq. 2 predicts whether each object is stationary or in motion, while eq. 3 predicts their pose. During testing, the output from the sigmoid is rounded to the nearest integer to emulate a binary output. As such, the final predicted object pose $\hat{x}_i^t$ would either be the output of φ_p or its pose at the previous timestep, depending on its binary probability score.

Human Pose Ensemble Module (PEM): Existing works that forecast (Corona et al. 2020) or synthesize (Razali and Demiris 2023) the human and object motion from scratch are limited in that they do not encourage hand-object contact during manipulation. Although the features are extracted jointly through a graph network, in reality, the prediction of the human is only weakly conditioned on the object and vice-versa. The consequence of this is objects that float in space that do not move together in sync with the hands. We address this by refining the predicted human pose with a method that encourages contact. We first predict the object in contact with the masked hand:

$$\hat{\beta}^t = \text{GS}(c^t), \text{where } c_j^t = \varphi_{c_2}([g_j^t, l_j, s]) \quad (5)$$

where GS is the Gumbel-Softmax estimator (Jang, Gu, and Poole 2017), the variable $s = [1, 0]$ or $[0, 1]$ is a one-hot vector for the left or right side of the arm that was masked, c_j^t is a scalar logit and c^t its corresponding vector. We feed the vector c^t to the Gumbel-Softmax estimator to return a one-hot vector $\hat{\beta}^t$ where each element $\hat{\beta}_j^t$ can be interpreted as a binary probability score. Note that, unlike φ_{c_1} which outputs an individual probability score for each object, φ_{c_2} outputs a distribution over the list of objects and human. This is done to account for when the masked hand is not manipulating an object i.e. $\hat{\beta}_h^t = 1$ and $\hat{\beta}_{1:n}^t = \vec{0}$ when there is no contact.

We then use an MLP Ψ to reconstruct the human pose and a GrabNet Φ (Taheri et al. 2020) to encourage contact if it exists. GrabNet is a conditional VAE that reconstructs the hand, conditioned on the Basis Point Set (BPS) (Prokudin, Lassner, and Romero 2019) representation of the object mesh. The BPS is an efficient non-learning method that samples a point cloud of the object mesh into a fixed-length representation. Then, because we have multiple objects in the scene, we run GrabNet for every object and use $\hat{\beta}_{1:n}^t$ to refine the reconstructed pose. We modify GrabNet in our implementation to a deterministic version to additionally refine the arms due to the variability of the arm pose based on the object's location relative to the person's body:

$$\hat{y}_\Psi^t = \Psi(g_h^t) \quad (6)$$

$$\hat{y}_{\Phi,i}^t = \Phi([y^t \circ m, x_i^t, \text{BPS}_i, s]) \quad (7)$$

$$\hat{y}^t = \underbrace{y^t \circ m_s}_{\text{w/o arm}} + \underbrace{\textstyle\sum_i (\hat{\beta}_i^t \hat{y}_{\Phi,i}^t}_{\text{arm w/ contact}} + \underbrace{\hat{\beta}_h^t \hat{y}_\Psi^t)}_{\text{arm w/o contact}} \circ (1 - m_s) \quad (8)$$

where $\circ$ denotes the elementwise multiplication. m_s and m are binary masks for the human joints where the subscript s denotes the side of the arm. Specifically, we set m_s to 0 from the shoulder to the fingers along the side of the arm that was masked i.e., $\bar{y}^t = y^t \circ m_s$. m then masks both sides of the arm. Our GrabNet in eq. 7 thus receives as input the human pose without both arms to predict the side that was masked, conditioned on the side of the arm s, object pose x_i^t, and its BPS. The mask m_s is then used to refine predictions for said arm in eq. 8 if there is contact. Likewise, $\hat{\beta}_h^t = 1$ and $\hat{\beta}_{1:n}^t = \vec{0}$ if there is no contact. In this case, the predicted arm would be the unrefined version from the MLP Ψ.

The primary difference between our novel use of the GrabNet and the MLP used in the final layer for decoding the human pose in prior works (Corona et al. 2020; Cui and Sun 2021; Razali and Demiris 2023) is that the former is

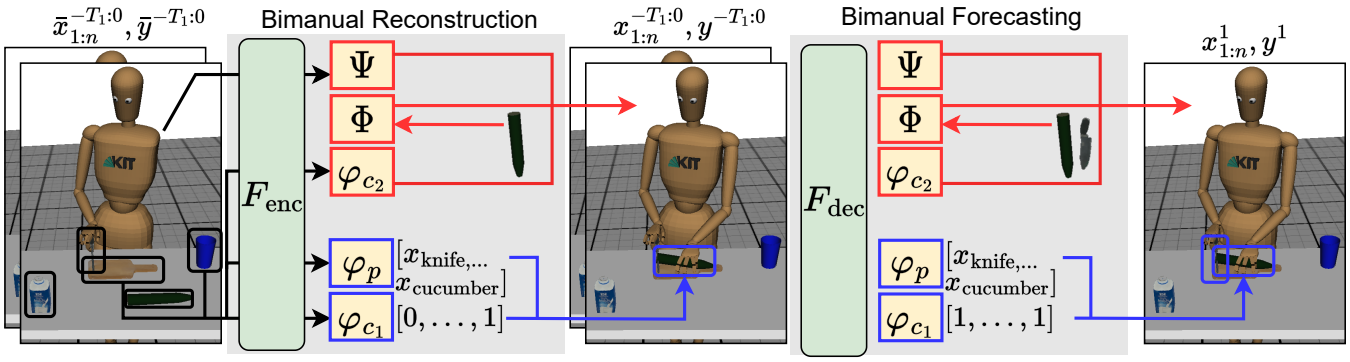

Figure 2: The pathways in blue form the object motion module (OMM) and in red the human pose ensemble module (PEM). During reconstruction, the OMM predicts the pose for the objects undergoing motion or manipulation, while the PEM reconstructs and refines the missing arm conditioned on the object in contact. During forecasting, the PEM refines the pose of each arm given the objects in contact. F_{enc} and F_{dec} denote Graph GRUs, φ and Ψ MLPs, and Φ a GrabNet.

explicitly conditioned on the human pose and object data while the latter only on the output of the graph network. The difference becomes further exacerbated especially since the decoders in said works were trained to minimize the pose at every timestep, whether there is contact or not. By contrast, the error signals in our work are partitioned: the GrabNet is trained explicitly during contact and the MLP at all times. The network then selects the decoder given the contact probabilities. This eases training and ensures that each model is trained to perform its respective task optimally.

Bimanual Forecasting

We use the OMM and PEM described earlier for forecasting but with slight modifications and a separate set of weights since there are now a maximum of two objects that can each be manipulated by the left and right hands. Consequently, the pose may need to be updated for both arms. To begin, we first compute g for only the next timestep, conditioned on the reconstructed past:

$$g_j^t = F_{\text{dec}} = \text{GRU}(\max_{j' \in \mathcal{N}(j)} \varphi([v_j^t, v_{j'}^t - v_j^t]); g_j^{t-1}) \quad (9)$$

where t is now $[0, ..., T_2]$. We then forecast the object pose velocities with another OMM:

$$\hat{\alpha}_i^t = \sigma(\varphi_{c_1}([g_i^t, l_i])) \quad (10)$$

$$\hat{x}_{\varphi,i}^t = \varphi_p([g_i^t, l_i]) + \hat{x}_i^{t-1} \quad (11)$$

$$\hat{x}_i^t = \hat{\alpha}_i^t \hat{x}_{\varphi,i}^t + (1 - \hat{\alpha}_i^t) x_i^{t-1} \quad (12)$$

We next forecast the human pose velocities then refine both arms, indexed again by the subscript $s = [l, r]$ for the left and right arms, respectively:

$$\hat{\beta}_s^t = \text{GS}(c_s^t), \text{where } c_{s,j}^t = \varphi_{c_2}([g_j^t, l_j, s]) \quad (13)$$

$$\hat{y}_\Psi^t = \Psi(g_{\text{h}}^t) + \hat{y}^{t-1} \quad (14)$$

$$\hat{y}_{\Phi,s,i}^t = \Phi([\hat{y}_\Psi^t \circ m, x_i^t, \text{BPS}_i, s]) \quad (15)$$

$$\hat{y}'^t = \hat{y}_\Psi^t \circ m_l + \textstyle\sum_i (\hat{\beta}_{l,i}^t \hat{y}_{\Phi,l,i}^t + \hat{\beta}_{l,\text{h}}^t \hat{y}_\Psi^t) \circ (1 - m_l) \quad (16)$$

$$\hat{y}^t = \hat{y}'^t \circ m_r + \textstyle\sum_i (\hat{\beta}_{r,i}^t \hat{y}_{\Phi,r,i}^t + \hat{\beta}_{r,\text{h}}^t \hat{y}'^t) \circ (1 - m_r) \quad (17)$$

Intuitively, eq. 15 clones the predicted pose twice then masks them as input to the GrabNet. The left and right arms are then sequentially replaced with the refined ones in eqs. 16 and 17 respectively if there is hand-object contact. Finally, the predicted human and object poses are fed back to eq. 9 for the next iteration. Note that we follow past related work to assume motion in a short future to be mostly deterministic (Corona et al. 2020; Cui and Sun 2021) although it is possible to inject stochasticity into the various components e.g., graph GRU with a VAE.

Training: We optimize for the initial predicted human pose $\hat{y}_\Psi$, object pose $\hat{x}_{\varphi,i}$, object motion $\hat{\alpha}_i$ and contact $\hat{\beta}_{s,j}$ probabilities at every timestep, and the arm component of $\hat{y}_{\Phi,s}$ but only when there is hand-object contact. Our architecture is then trained end-to-end with the following objective:

$$\sum_{s,i,j,t} \lambda_1 (\underbrace{\|\hat{y}_\Psi^t - y^t\|_2^2}_{\text{human pose}} + \underbrace{\|\hat{x}_{\varphi,i}^t - x_i^t\|_2^2}_{\text{object pose}}) + \lambda_2 \underbrace{\hat{\alpha}_i^t \log \alpha_i^t}_{\text{object motion probability}}$$
$$+ \lambda_1 \underbrace{\delta_s^t \|(\hat{y}_{\Phi,s}^t - y^t) \circ (1 - m_s)\|_2^2}_{\text{refined arms}} + \lambda_2 \underbrace{\hat{\beta}_{s,j}^t \log \beta_{s,j}^t}_{\text{contact probability}}$$

where δ is 1 during contact and 0 otherwise, and $\{\lambda_1, \lambda_2\} = \{10^{-2}, 10^{-1}\}$. The gradients for the contact probabilities are computed via the straight-through trick (Jang, Gu, and Poole 2017).

Experiments

The **KIT Bimanual Manipulation (KIT MoCap)** (Krebs et al. 2021) dataset contains 2 hour motion capture recordings of right-handed individuals performing everyday bimanual tasks in the kitchen such as cutting a fruit or pouring water into a cup. The dataset contains a mesh of each object whose pose is inferred via 4 motion capture markers on the object. The dataset has been manually annotated to provide temporal contact information but lacks vertex-level hand-object contact (Taheri et al. 2020). As such, the GrabNet in

our implementation only consists of the CoarseNet module. For each sequence, we sample various objects from different sequences, employing them as distractors with a maximum limit of four. We follow the preprocessing described in (Razali and Demiris 2023) to add noise and center the coordinates with respect to the table center. We train our model to observe the past 10 time steps (1 second) to predict the future 20 (2 seconds). Lastly, we have the variable $x_i^t \in R^{K\times 3}$ denote the 4 markers on the object i.e. $K = 4$. The **KIT RGBD Bimanual Actions (KIT RGBD)** (Dreher, Wächter, and Asfour 2019) dataset also contain 2 hour recordings of bimanual tasks in the kitchen but is recorded using an RGBD camera. We use the provided human pose and object bounding boxes that are estimated using OpenPose (Cao et al. 2017) and YOLOv3 (Redmon and Farhadi 2018) respectively. As the dataset does not contain temporal contact information, we estimate it using proximity between the 3D hand and object bounding boxes, and the manually annotated temporal bimanual action labels. Specifically, contact can only be made during actions such as "Hold" or "Cut". During such actions, the hand-object pair with the smallest distance is assumed to be in contact. We also freeze the motion for objects not in contact with either hands as using the raw detections from YOLO would otherwise introduce drift in our ground truth. As the dataset is significantly more challenging due to noisy detections exacerbated by imperfect point clouds and temporal contact annotations, we make the following changes to our model. First, we exclude the finger pose from GrabNet to have it refine only the arm during contact. Second, we represent the object by its centroids as we find the 3D bounding boxes to be highly noisy. In this case, $x_i^t \in R^{K\times 3}$ denotes the object center i.e. $K = 1$.

Experimental Setup

We adopt the widely utilized Mean Per Joint Position Error (Corona et al. 2020; Cui and Sun 2021) as a metric to assess the models' performance. We mask each arm once and report separate human and object pose errors, with object errors categorized into distractor for objects not held by the hand, and contact for those undergoing manipulation. The overall error is the weighted average of these categories. We compare our approach to the most closely related existing methods: (i) The context-aware human motion prediction (CAHMP) (Corona et al. 2020) (ii) the multi-task graph convolutional network (MTGCN) (Cui and Sun 2021) that we modify to include the object label and pose, and (iii) the modularized architecture (Mod) (Razali and Demiris 2023) for synthesis that we adapt for forecasting given an input sequence. To ensure fairness, we train both Mod and CAHMP to reconstruct the bimanual sequence during the encoding phase before forecasting by predicting the missing arm and every object in the scene, except for the object observed to be undergoing manipulation. MTGCN in contrast, already has a built-in reconstruction module. Furthermore, as Mod is a stochastic model, we average its output over 16 runs. We keep our model deterministic using off-the-shelf components for fairness and to further emphasise our contributions. All models contain 3 million parameters, and reconstruct and forecast the pose positions and velocities respectively. Lastly, the experiments were implemented in PyTorch and trained using the ADAM (Kingma and Ba 2015) optimizer for 500 epochs with a batch size of 64 until convergence, for about 12 hours on a NVIDIA RTX 2080.

Quantitative Results

We present results on the KIT MoCap and KIT RGBD datasets in Tables 1 and 2 respectively. The columns -1:0 and 0:2 indicate the results for reconstruction and forecasting respectively, and the column 0:2* for forecasting with a bimanual ground truth input. We also include the contact classification accuracies for our method. Our approach outperforms the state-of-the-art (SOTA) model (Mod) for both reconstruction and forecasting by a significant margin. Most notably, the SOTA shows high distractor errors which in turn negatively affects the overall errors. Ours on the other hand, have very low distractor errors. We attribute this to our OMM that suppresses predictions for objects not in motion. It is far easier to suppress the distractors through the aid of a classifier that outputs binary probabilities than it is to have a network output exact zeros for the pose. The PEM's ability to identify objects in contact also translate to lower human pose errors due to refinements made by the GrabNet. These naturally rely on the classifier being performant enough though it can be inferred that our obtained accuracies provide a net benefit. Note that the contact errors are inherently higher as they are averaged for objects undergoing manipulation. While our numbers degrade on the more challenging KIT RGBD dataset, we still outperform the SOTA.

As an ablation study, we assessed our models in two conditions: one without the proposed components (w/o OMM + PEM) — meaning no motion suppression nor pose refinements — and the other with these components, but employing the ground truth object motion labels and contact labels (w/ GT). The results demonstrate that the absence of our components leads to a degradation in model performance, resulting in errors worse than the SOTA. In contrast, utilizing the ground truth labels yields a further performance boost. Overall, these findings provide compelling evidence for the efficacy of our components in forecasting bimanual manipulation sequences.

Qualitative Results

We present a comparison on the KIT MoCap dataset in Figure 3, where the input is a unimanual sequence that has the right arm (in red), and respective object (peeler) masked. We compare our output to Mod and additionally provide visualizations of our contact classifications for both hands. Frames 1 and 2 show reconstruction and 3 to 8 forecasting. Recall that the masked object does not move during the observed unimanual interaction and may need to have its pose repredicted by the model during bimanual reconstruction.

Our model accurately reconstructs the bimanual sequence, with the right hand appropriately interacting with the peeler instead of other objects due to the left-hand approach to the cucumber. The forecasting is also accurate, with both hands maintaining contact with their respective objects while the right hand peels the cucumber held by the left. The roller and cutting board, which are not manipulated,

	Distractor (mm)			Contact (mm)			Overall (mm)			Accuracy (%)			Human (mm)		
Method	−1:0	0:2	0:2*	−1:0	0:2	0:2*	−1:0	0:2	0:2*	−1:0	0:2	0:2*	−1:0	0:2	0:2*
MTGCN	14.9	33.4	31.0	57.2	110.2	102.1	21.9	54.6	50.3	-	-	-	25.2	60.1	55.6
CAHMP	13.5	29.4	25.8	54.5	107.1	99.3	20.1	52.4	47.7	-	-	-	23.9	58.7	52.9
Mod	12.4	27.5	22.1	53.8	106.2	99.1	18.5	51.3	44.4	-	-	-	21.9	58.2	51.2
Ours	**0.3**	**4.9**	**3.4**	**50.7**	**100.2**	**92.0**	**8.9**	**33.7**	**30.3**	95.1	90.1	92.4	**18.3**	**54.2**	**46.6**
w/o OMM + PEM	13.7	29.3	25.9	54.8	107.6	99.5	20.3	52.7	48.1	-	-	-	23.8	58.8	52.9
w/ GT	0.0	0.0	0.0	48.8	97.2	90.4	8.8	32.1	28.9	100	100	100	17.2	52.7	45.4

Table 1: Results on the KIT MoCap dataset. Our method surpasses the SOTA in both human and object pose reconstrution and forecasting with significantly lower distractor errors.

	Distractor (cm)			Contact (cm)			Overall (cm)			Accuracy (%)			Human (cm)		
Method	−1:0	0:2	0:2*	−1:0	0:2	0:2*	−1:0	0:2	0:2*	−1:0	0:2	0:2*	−1:0	0:2	0:2*
MTGCN	23.2	35.7	35.1	41.8	62.9	61.1	22.5	50.1	49.3	-	-	-	25.8	41.22	39.2
CAHMP	19.3	33.2	32.4	38.2	59.4	58.0	19.6	45.7	44.2	-	-	-	21.3	36.1	35.5
Mod	18.8	32.8	31.9	36.4	56.6	54.6	17.7	44.3	42.1	-	-	-	20.2	35.6	34.6
Ours	**3.39**	**10.8**	**10.2**	**21.0**	**45.0**	**42.2**	**8.7**	**25.8**	**24.4**	86.5	84.0	85.6	**16.4**	**31.9**	**29.8**
w/o OMM + PEM	20.2	35.4	34.3	39.5	61.5	60.4	20.9	46.3	44.8	-	-	-	22.2	36.7	36.2
w/ GT	0.0	0.0	0.0	18.3	40.4	38.1	10.3	24.1	23.1	100	100	100	15.8	27.1	25.8

Table 2: Our method continues to show improvements over the SOTA on the noisier KIT RGBD dataset.

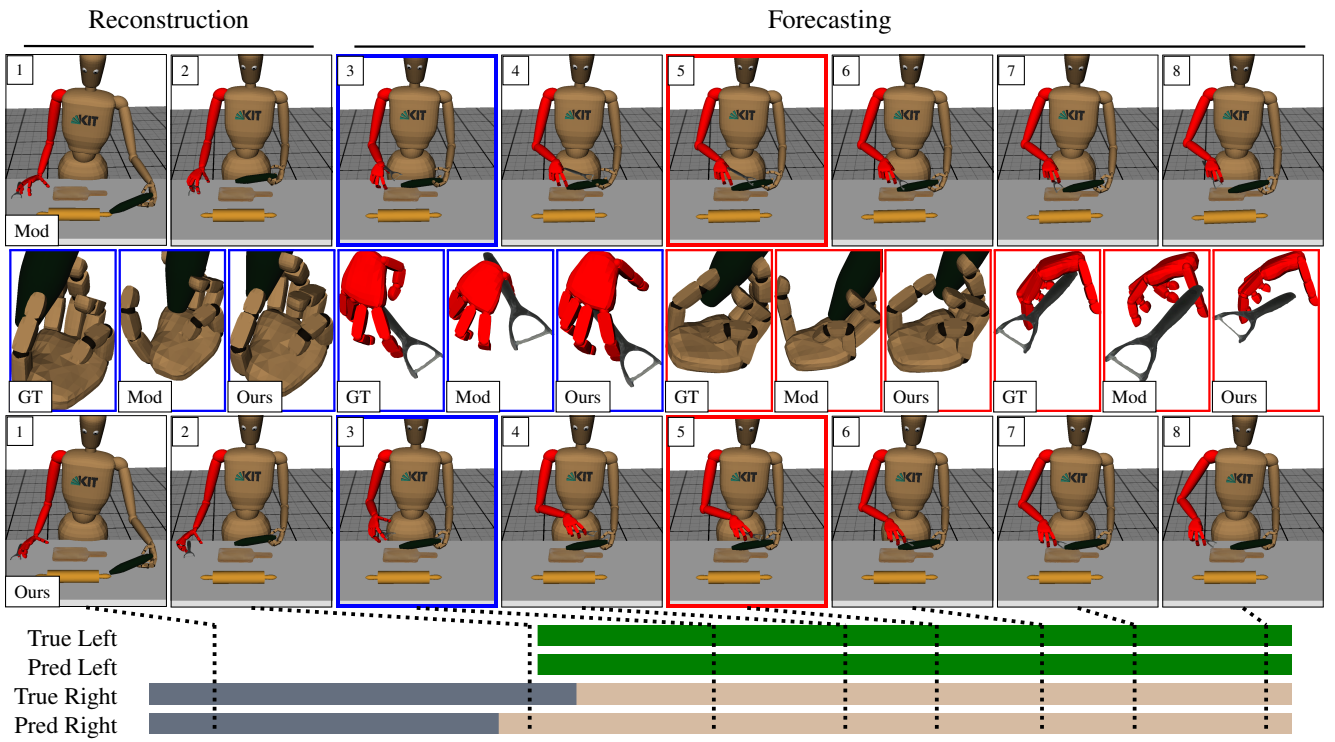

Figure 3: KIT MoCap results with the right arm and peeler masked. Frames 1 and 2 show reconstruction, and 3 to 8 forecasting. Our results show better hand-object contact for the object (peeler) in motion, especially from frames 3 to 8. The hand pose is shown for frames 3 and 5, with the frame borders colored in blue and red respectively. The colored bars show the true and predicted object in contact for the left and right hands, with the lines connecting the respective timesteps. Classification is not done for the observed left during reconstruction. Cucumber Peeler No contact.

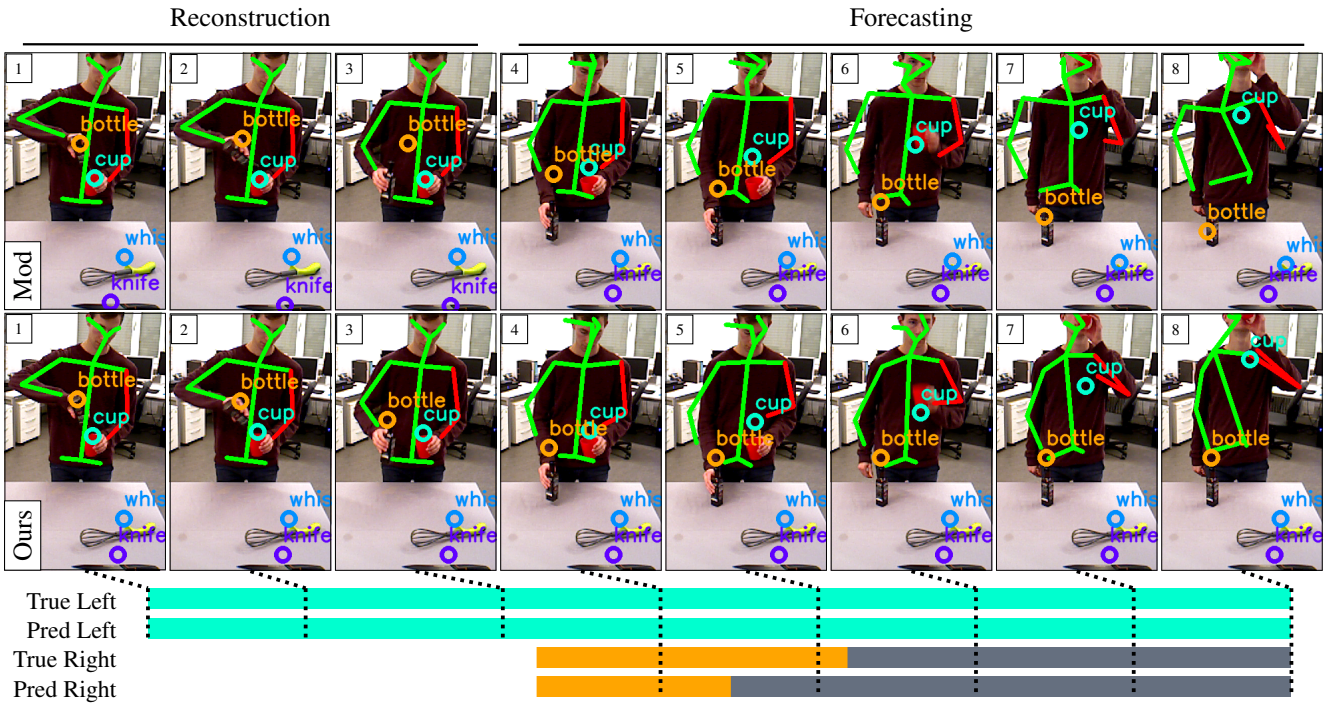

Figure 4: KIT RGBD results with the left arm and cup masked. Frames 1 to 3 depict reconstruction, while the rest forecasting. Our method shows better human pose forecasting and visible improvements in encouraging contact between the left wrist and the cup. Our distractors also remain motionless while the SOTA exhibits some drift. Cup Bottle No contact.

remain fixed in place. These improvements are achieved due to our OMM and PEM. The OMM's classifier suppresses objects not in motion while the PEM utilizes the GrabNet to refine the human pose in relation to the objects in contact. The classification bars also show that our model precisely identifies the object each hand is in contact with. This observation is crucial as it indicates the GrabNet being engaged at the correct timesteps, albeit with errors primarily occurring at the moment of contact. Note that the figure does not include the classification for the observed left during reconstruction. In contrast, the motion generated by the SOTA is less desirable due to the peeler's incorrect position relative to the right hand. Observe specifically how in frames 3 to 8 the peeler continues to move along the cucumber, away from the hand while the hand remains almost stationary. The objects not undergoing manipulation also exhibit some drift, more clearly seen in the video supplementary. The undesired motion of the hand relative to the peeler is primarily caused by the MLP decoder, which is not explicitly conditioned on the object in contact and is trained to minimize errors at every timestep, regardless of contact. On the other hand, the drift observed in non-manipulated objects can be attributed to the MLP's inability to output precise zero values for the pose. Lastly, we observe some gap in the hand-object contact from our model, although this is due to the absence of vertex-level annotations in the dataset as shown in the ground truth.

We present additional results on the challenging KIT RGBD dataset in Figure 4, where the left arm and cup have been masked, and the points reprojected onto the image. Our model outperforms the SOTA by demonstrating improvements in the placement of especially the left wrist in relation to the cup it is holding. Frames 6 to 8 also show our method exhibiting more stable and less noisy outputs than the SOTA. We credit this to our OMM that suppresses the outputs for objects not undergoing manipulation. This approach reduces noise in the inputs during the recurrent forecasting stage, leading to more stable and improved outputs.

Conclusion

We proposed a novel method for forecasting bimanual manipulation sequences. Our method predicts the motion for only the objects undergoing manipulation before refining the human pose conditioned on the objects in contact. As a result, our method yields more plausible and accurate human and object motions than the SOTA that has more sophisticated graph architectures. The components are also generic enough for use in both unimanual and bimanual contexts, allowing us to forecast bimanual sequences from unimanual observations. Our experiments are limited as the sequence was generated by masking the data which may not be reflective of the motion exhibited by a person with one functional arm. An individual with such a disability may, over time, modify their strategies when carrying out tasks typically performed with two hands. A better approach would be to collect recordings of individuals with only one functional hand, despite the challenge of this task.

References

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Corona, E.; Pumarola, A.; Alenya, G.; and Moreno-Noguer, F. 2020. Context-aware human motion prediction. In *Proceedings of the 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4801–4810.

Dreher, C. R.; Wächter, M.; and Asfour, T. 2019. Learning object-action relations from bimanual human demonstration using graph networks. *IEEE Robotics and Automation Letters*, 5(1): 187–194.

García, A.; Solanes, J. E.; Muñoz, A.; Gracia, L.; and Tornero, J. 2022. Augmented Reality-Based Interface for Bimanual Robot Teleoperation. *Applied Sciences*, 12(9): 4379.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hassan, M.; Ghosh, P.; Tesch, J.; Tzionas, D.; and Black, M. J. 2021. Populating 3D scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14708–14718.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. MotionGPT: Human Motion as a Foreign Language. *arXiv preprint arXiv:2306.14795*.

Johansson, R. S.; and Cole, K. J. 1992. Sensory-motor coordination during grasping and manipulative actions. *Current opinion in neurobiology*, 2(6): 815–823.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koppula, H. S.; and Saxena, A. 2013. Anticipating human activities for reactive robotic response. In *IROS*, 2071. Tokyo.

Krebs, F.; Meixner, A.; Patzer, I.; and Asfour, T. 2021. The KIT Bimanual Manipulation Dataset. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, 499–506. IEEE.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13401–13412.

Liu, Z.; Lyu, K.; Wu, S.; Chen, H.; Hao, Y.; and Ji, S. 2021. Aggregated multi-gans for controlled 3d human motion prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2225–2232.

Lum, P. S.; Burgar, C. G.; Shor, P. C.; Majmundar, M.; and Van der Loos, M. 2002. Robot-assisted movement training compared with conventional therapy techniques for the rehabilitation of upper-limb motor function after stroke. *Archives of physical medicine and rehabilitation*, 83(7): 952–959.

Martinez, J.; Black, M. J.; and Romero, J. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2891–2900.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10985–10995.

Prokudin, S.; Lassner, C.; and Romero, J. 2019. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4332–4341.

Razali, H.; and Demiris, Y. 2022a. Using a Single Input to Forecast Human Action Keystates in Everyday Pick and Place Actions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3488–3492. IEEE.

Razali, H.; and Demiris, Y. 2022b. Using Eye Gaze to Forecast Human Pose in Everyday Pick and Place Actions. In *2022 International Conference on Robotics and Automation (ICRA)*, 8497–8503. IEEE.

Razali, H.; and Demiris, Y. 2023. Action-Conditioned Generation of Bimanual Object Manipulation Sequences. In *Proceedings of the AAAI conference on artificial intelligence*.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Starke, S.; Zhang, H.; Komura, T.; and Saito, J. 2019. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6): 209–1.

Taheri, O.; Choutas, V.; Black, M. J.; and Tzionas, D. 2022. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13263–13273.

Taheri, O.; Ghorbani, N.; Black, M. J.; and Tzionas, D. 2020. GRAB: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 581–600. Springer.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*.

Zheng, Y.; Yang, Y.; Mo, K.; Li, J.; Yu, T.; Liu, Y.; Liu, C. K.; and Guibas, L. J. 2022. Gimo: Gaze-informed human motion prediction in context. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, 676–694. Springer.