COUCH: Towards Controllable Human-Chair Interactions

Xiaohan Zhang[1,2], Bharat Lal Bhatnagar[1,2], Vladimir Guzov[1,2], Sebastian Starke[3,4], and Gerard Pons-Moll[1,2]

[1] University of Tübingen, Germany
[2] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[3] Electronic Arts
[4] University of Edinburgh, United Kingdom

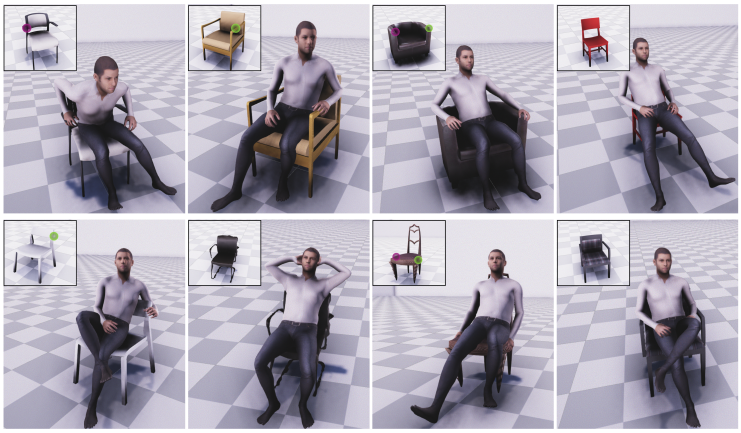

Fig. 1: We present COUCH. A dataset and model to synthesizes controllable, contact-based human-chair interactions.

Abstract. Humans interact with an object in many different ways by making contact at different locations, creating a highly complex motion space that can be difficult to learn, particularly when synthesizing such human interactions in a controllable manner. Existing works on synthesizing human scene interaction focus on the high-level control of action but do not consider the fine-grained control of motion. In this work, we study the problem of synthesizing scene interactions conditioned on different contact positions on the object. As a testbed to investigate this new problem, we focus on human-chair interaction as one of the most common actions which exhibit large variability in terms of contacts. We propose a novel synthesis framework COUCH that plans ahead the motion by predicting contact-aware control signals of the hands, which are then used to synthesize contact-conditioned interactions. Furthermore, we contribute a large human-chair interaction dataset with clean annotations, the COUCH Dataset. Our method shows significant quantitative and qualitative improvements over existing methods for human-object interactions. More importantly, our method enables control of the motion through user-specified or automatically predicted contacts. Our dataset, model and code will be available at [1].

1 Introduction

To synthesize realistic virtual humans which can achieve goals and act upon the environment, reasoning about the interactions and in turn contacts, is necessary. Reaching a goal, like sitting on a chair, is often preceded by intentional contact with the hand to support the body. In this work we investigate a motion synthesis method which exploits predictable contact to achieve more control and diversity over the animations.

Although most applications in VR/AR, digital content creation and robotics require synthesizing motion *within the environment*, it is not considered in the majority of works in human motion synthesis [40,41,39,33]. Recent work does take the environment into account but is limited to synthesizing *static poses* [23,58]. Synthesising dynamic human motion coherent with the environment is a substantially harder task [47,22,26,48,49] and recent works show promising results. However, these methods do not reason about intentional contacts with the environment, and can not be controlled with user provided contacts.

Thus, in this work, we investigate a new problem: synthesizing human motion conditioned on contact positions on the object to allow for controllable movement variations. As a testbed to investigate this new problem, we focus on human-chair interactions as one of the most common actions, which are of crucial importance for ergonomics, avatars in virtual reality or video game animation. Contact-driven motion synthesis is a more challenging learning problem compared to conditioning only on coarse object geometry [47,22]. First, the human needs to approach the chair differently depending on the contacts, regardless of the starting position, walking around it if necessary. Second, a chair can be approached and contacted in many different ways; we can directly sit without using our hands, or we can first support the body weight using the left/right or both hands with different parts of the chair. Furthermore, individual styled free-interactions can be modelled such as leaning back, stretching legs, using hands to support the head, and so on.

Contact driven motion allows for providing more detailed instructions to the virtual human such as approaching to sit on the chair, while supporting the body with the left hand and placing it on the armrest, as illustrated in Figure 1. Given the contact and the goal, the full-body needs to coordinate at run-time to achieve a plausible sequence of pose transitions. Intuitively, this emulates our planning of motion as real humans: we plan in terms of goals and intermediate object contacts to reach; the full-body then moves to follow such desired trajectories.

To this end, we propose COUCH, a method for controllable contact driven human-chair interactions, which is composed of two core components: 1) *ControlNet* is responsible for motion planning by predicting the future control signal of the body limbs which guides the future body movement. Our spatial-temporal control signal consists of dynamic trajectories of the hands towards the contact points and the local phase, an auxiliary continuous variable that encodes the temporal information of a limb during a particular movement (e.g. the left hand reaching an armrest). 2) *PoseNet* conditions on the predicted control signal to synthesise motion that follows the dynamic trajectories, ensuring the con-

tact point is reached. At runtime, COUCH can operate in two modes. First, in an interactive mode where the user specifies the desired contact points on the target object. Second, in a generative mode where COUCH can automatically sample diverse intentional contacts on the object with a novel neural network called *ContactNet*. Training and evaluating COUCH calls for a dataset of rich and accurate human chair interactions. Existing interaction 3D datasets [47] are captured with Inertial Sensors, and hence do not capture the real body motion and the contacts with the real chair geometry – instead synthetic chairs are fit to the avatar as post-process in those works. Hence, to jointly capture real human-chair interactions with fine-grained contacts, we fit the SMPL model [37] and scanned chair models to data obtained from multiple Kinects and IMUs. The dataset (the COUCH dataset, Table 4) consists of 3 hours (over 500 sequences) of motion capture (MoCap) on human-chair interactions. Moreover it features multiple subjects, accurately captured contacts with registered chairs, and annotation on the type of hand contact. Our experiments demonstrate that COUCH runs in real-time at 30 fps, COUCH generalizes across chairs of varied geometry, and different starting positions relative to the chair. Compared to SoTA models (trained on the same data) adapted to incorporate contacts, our method significantly outperforms them in terms of control by improving the average contact distance by 55%.

The contributions of our work can be summarized as follows:

- We propose COUCH, the first method for synthesizing controllable contact-based human-chair interactions. Given the same input control, the COUCH model can achieve diverse sitting motions. By specifying different control signals, the user enables control over the style of interaction with the object. Results show our method outperforms the state of the art both qualitatively and quantitatively.
- To train COUCH, we captured a large-scale MoCap dataset consisting of 3 hours (over 500 sequences) of human interacting with chairs different styles of sitting and free movements. The dataset features multiple subject, real chair geometry, accurately annotated hand contacts, and RGB-D images.
- To stimulate further research in controllable synthesis of human motion, we will release the COUCH model and dataset.

2 Related Work

Scene agnostic human motion prediction. Synthesizing realistic human motion has drawn much attention from the computer vision and graphics communities. However, many methods do not take the scene into account. Existing methods on short (∼1 sec) [40,20,41,16,54,51,5,55,24] and long (>1 min) [28,18] term 3D human motion prediction aim to produce realistic-looking human motion (typically walking and its variants). There also exists work on conditional motion generation based on music [33]. These methods have two major limitations, i) except for work that use generative models [36,21,19,6,25], these methods

are largely deterministic and cannot be used to generate diverse motions and ii) this body of work is unfortunately agnostic to scene geometry [32,39], which is critical to model human scene interactions. Our method on the other hand can generate *realistic motion and interactions*, taking into account the 3D scene.

Affordance and Static Scene Interactions. Although the focus of our work is to model human-scene interactions over time, we find the works predicting static affordances in a 3D scene [34] relevant. This branch of work aims at predicting static humans in a scene [58,57,23] that satisfies the scene constraints. More recently there have been attempts to model fine-grained interactions (contacts) between the hand and objects [15,50,9,31].

The aforementioned methods focus on predicting static humans / human poses that satisfy the scene constraints in case of affordances or grasping an object in case of hand-object interactions. But these methods cannot produce a full sequence of human motion and interaction with the scene. Ours is the first approach that can model fine-grained interactions (contacts) between an object in the scene and the human.

Dynamic Scene Interactions. Although various algorithms have been proposed for scene-agnostic motion prediction, affordance prediction as well as the synthesis of static human-scene interaction, generating dynamic human-scene interactions is less explored. Recent advances include predicting human motion from scene images [10], and using a semantic graph and RNN to predict human and object movements [14]. More recently, Wang et al. [52] introduce a hierarchical framework that generates 'in-between' locations and poses on the scene and interpolates between the goal poses. However, it requires a carefully tuned post-optimization step over the full motion synthesis to solve the discontinuity of motion between sub-goals and to achieve robust foot contacts with the scene. Alternatively, Chao et al. [13] use a reinforcement learning based approach by training a meta controller to coordinate sub-controllers to complete a sitting task. An important category of human-scene interaction involves performing locomotion on uneven terrains. The Phase-functioned Neural Network [27] first introduced the use of external phase variables to represent the state of the motion cycle. Zhang et al. [56] applies same concept for quadruped motion and further incorporates a gating network that segments the locomotion modes based on foot velocities. Both works show impressive results thanks to the mixture of experts [17] styled architectures.

The most relevant work to us, are the Neural State Machine (NSM) [47] and SAMP [22]. While NSM is a powerful method and models human-scene interactions such as sitting, carrying boxes and opening doors, it does not generate motion variations for the same task and object geometry. SAMP predicts diverse goal locations in the scene for the virtual human, which is then used to drive the motion generation. Our work takes inspiration from these works, but it is demonstrated qualitatively and quantitatively from our experiments that neither of the work enables control over the style of interaction (Section 5.2). Our

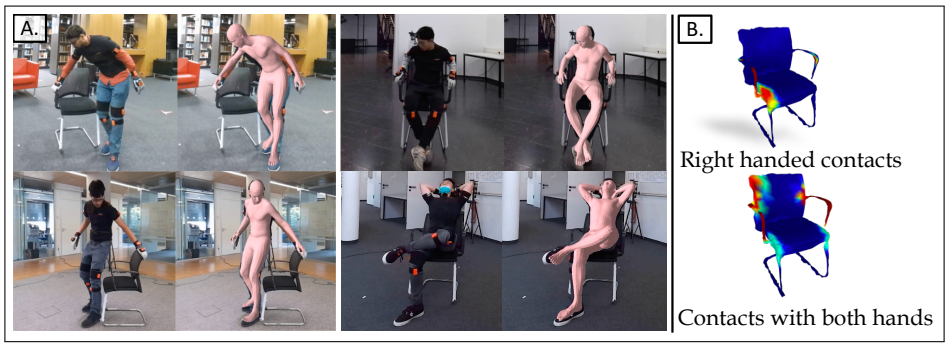

Fig. 2: (A) The COUCH dataset captures a diverse range chair interaction with an emphasis on hand contacts. It consists of RGB-D images, corresponding MoCap in the SMPL [37] format, chair geometry and annotations on types of hand interaction. (B) COUCH dataset captures natural modes of interactions with a chair, as demonstrated by the heatmaps of contact clusters. Most common contacts while sitting, include right hand support or both hands.

work focus on controllable, fine-grained interactions given on contacts on the object. To the best of our knowledge, no previous work has tackled the problem of generating controllable human-chair interactions.

3 The COUCH Dataset

Large scale datasets of human motion such as AMASS [38] and H3.6m [29] have allowed us to build models of human motion. Unfortunately these datasets only contain sequences of 3D poses but no information about the 3D environment, making these datasets unsuitable for learning human interactions. On the other hand, datasets containing human-object interactions are either restricted to just hands [9], contain only static human poses [50,23] without any motion, or have little variation in motion [47].

We present a multi-subject dataset of human chair interactions. Our dataset consists of 6 different subjects interacting with chairs with over 500 motion sequences. We collect our dataset using 17 wearable Inertial Measurement Units (XSens) [4], from which we obtain high-quality pose sequences in SMPL [37] format using Unity [30]. The total capture length is 3 hours.

Motion capture with marker-based capture systems is restrictive to capturing human-object interactions because markers often get occluded during the interactions leading to inaccurate tracking. IMU-based systems are prevalent for large-scale motion capture, however, the error from its calibration can lower the accuracy of the motion. We propose to combine IMUs with Kinect-based capture system as an efficient trade-off between scalability and accuracy. Our capture system is lightweight and can be generalized to capture many human

Table 1: Comparison with existing motion capture datasets for human chair interactions. The COUCH dataset features registered real chairs models, multiple subject, and RGB-D data. The types of hand contact are also annotated.

Features	NSM[47]	SAMP[22]	Ours
Real Objects	✗	✓	✓
Multiple Subjects	✗	✗	✓
Contact Types	✗	✗	✓
RGB-D	✗	✗	✓

scene interactions. We use the SMPL registration method similar to [8,44,7] to obtain SMPL fits for our data. The dataset is captured in four different indoor scenes. The average fitting error for the SMPL human model, and the chair scans to the point clouds from the Kinects are 3.12 cm and 1.70 cm, respectively (in Chamfer distance). More details about data capture can be found in supp. mat.

Diversity on Starting Points and Styles. We capture people approaching the chairs from different starting points surrounding the chairs. Each subject then performs different styles of interactions with the chairs during sitting. This includes, one hand touching the arm of the chair, both hands touching the armrests of the chair, one hand touching the sitting plane of the chair before sitting down, and no hand contacts. It also includes free interactions such as crossing legs or leaning forward and backward on the chairs. To ensure the naturalness of motion, each subject is only provided with high-level instruction before capturing each sequence and was asked to perform their styles freely. Annotations of the direction of the starting points relative to the chair as well as the type of hand contact are included in the dataset.

Objects. Our dataset contains three different chair models that vary in terms of their shapes, as well as a sofa. The objects are 3D scanned [3,2] before registering into the Kinect captured point clouds. To generalize the synthesized motion to unseen objects, we perform data augmentation as in [47].

Contacts. Studying contact-conditioned interaction calls for accurate contacts to be annotated in the dataset. Since we capture both the body motion and the object pose, it is possible to capture contacts between the body and the object. We detect the contacts of five key joints of the virtual human skeleton, which are the pelvis, hands, and feet. We then augment our data by randomly switching or scaling the object at each frame. The data augmentation is performed on 30 instances from ShapeNet [12] over categories of straight chairs, chairs with arms, and sofas. At every frame, we project the contacts detected from the ground truth data to the new object, and apply full-body inverse kinematics to recompute the pose such that the contacts are consistent, keeping the original context of the motion.

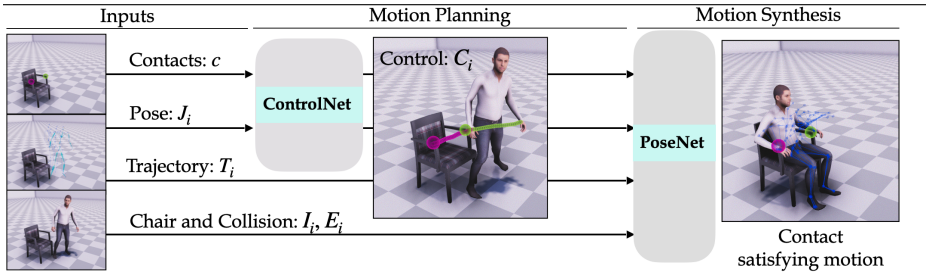

Fig. 3: Given user specified or model generated contacts, our proposed method which consists of the ControlNet and the PoseNet, auto-regressively synthesizes contact satisfying motion.

4 Method

We address the problem of synthesising 3D human motion that is natural and satisfies environmental geometry constraints and user-defined contacts with the chair. COUCH allows fine-grained control over how the human interacts with the chair. At run-time, our model operates in two modes. First, a generative mode where COUCH can automatically sample diverse intentional contacts on the object with our proposed generative model. Second, an interactive mode where the user specifies the desired contact points on the target object.
The input to our method is the current character pose, the target chair geometry as well the target contacts for the hands that need to be met. Our method takes these inputs and predicts the future poses that satisfy the desired contacts auto-regressively.

4.1 Key Insights

Synthesising natural human motion subject to environmental constraints is a challenging task [47,22,27], particularly when also satisfying a set of desired contacts. To this end, we first divide our motion synthesis task into *motion planning* and *motion prediction*. We derive our intuition from the way humans execute complex interactions e.g., to sit on the chair, we first prepare a mental model of how we will sit (place a hand on the arm-rest and sit, place a hand on the sitting plane and sit or just sit without using the hands etc.) and then we move our bodies accordingly. We propose two neural networks *ControlNet* $f^{\mathrm{CN}}(\cdot)$, and *PoseNet* $f^{\mathrm{PN}}(\cdot)$, for motion planning and motion prediction respectively.
Furthermore, we observe that it is useful to perform detailed hand motion planning only when we are close to the chair right before sitting and not when we are far off. Thus, we decompose the motion synthesis into *approaching* and *sitting*. The *approaching* motion can be generated directly with *PoseNet* but both networks are required for *sitting*, *ControlNet* and *PoseNet*, for generating the sitting motion that satisfies the given contacts.

4.2 Motion Planning with ControlNet

ControlNet is the core of our method and plays an important role in motion planning, that is predicting the future control signals of the key joints which are used to guide the body motions. At a high level, the contact-aware control signal contains the local phases and the future locations of the key joints (in our case, the two hands). The local phase is an auxiliary variable that encodes the temporal alignment of motion for each of the hands and prepares for a future contact to be made. When the virtual human is ready to make contact with the chair, and at the beginning of the hand movement, the local phase is equal to 0, and it gradually reaches the value 1 as the hand comes closer to the contact. The hand trajectory, on the other hand, encodes the spatial relationship between the hand joint and the given contact location.

More formally, we define our spatial-temporal control signal at frame $i+1$ to be $\mathbf{C}^+_{i+1} = \{\mathbf{h}^+_{i+1}, \phi^+_{i+1}\}$, where $\mathbf{h}^+_{i+1} \in \mathbb{R}^{2\times3\times\mathcal{T}^+}$ represents the future position of the two hand joints relative to their corresponding desired contact point $\boldsymbol{c} \in \mathbb{R}^{2\times3}$, and their local phases are represented by $\phi^+_i \in \mathbb{R}^{2\times\mathcal{T}^+}$. We predict the control signal for $\tau^+ = 7$ time stamps sampled uniformly between [0, 1] second window centered at frame $i+1$.
We use an LSTM based $f^{\mathrm{CN}}(\cdot)$ to predict the control signal,

$$\mathbf{C}^+_{i+1} = f^{\mathrm{CN}}(\tilde{\mathbf{h}}_i, \phi_i), \tag{1}$$

where $\tilde{\mathbf{h}}_i \in \mathbb{R}^{2\times3\times\mathcal{T}^+}$ denote τ^+ points interpolated uniformly on the straight line from the current hand locations to their desired contact locations $\boldsymbol{c}$. Intuitively, these interpolated positions encourages the ControlNet to predict future hand trajectories that always reach the given contacts. $\phi_i \in \mathbb{R}^{2\times\mathcal{T}}$ denotes the local phases of the hands over $\tau = 13$ frames sampled uniformly between the [-1, 1] second window centered at frame i.

The ControlNet is trained to minimize the following MSE loss on the future hand trajectories and the local phase, which is formulated as follows:

$$L_{\mathrm{control}} = \lambda_1\|\mathbf{h}^+_{i+1} - \hat{\mathbf{h}}^+_{i+1}\|^2_2 + \lambda_2\|\phi^+_{i+1} - \hat{\phi}^+_{i+1}\|^2_2 + \lambda_3 L_{\mathrm{reg}}. \tag{2}$$

Here, $\mathbf{h}^+_{i+1}, \phi^+_{i+1}$ are the network predicted future trajectories and local phases. $\hat{\mathbf{h}}^+_{i+1}, \hat{\phi}^+_{i+1}$ are the corresponding GT. We also introduce an additional regularization term $L_{\mathrm{reg}} = \|\mathbf{h}^+_{i+1} - \tilde{\mathbf{h}}_i\|_2$. Please see supplementary for implementation details regarding the network architectures and training.

4.3 Motion Synthesis with PoseNet

ControlNet generates important signals that guide the motion of the person such that user-defined contacts are satisfied. To this end, we train PoseNet $f^{\mathrm{PN}}(\cdot)$, that takes as input the control signals predicted by the ControlNet along with the 3D scene and motion in the past and predicts full body motion.

$$\mathbf{J}_{i+1}, \mathbf{T}^{+}_{i+1}, \mathbf{G}^{+}_{i+1}, \mathbf{\Phi}_{i+1}, \tilde{\boldsymbol{j}}^{p}_{i+1}, \tilde{\mathbf{T}}_{i+1}, \boldsymbol{b}_{i+1} = f^{\mathrm{PN}}(\mathbf{C}^{+}_{i}, \mathbf{J}_i, \mathbf{T}_i, \mathbf{G}_i, \mathbf{I}_i, \mathbf{E}_i, \mathbf{\Phi}_i), \quad (3)$$

where $\mathbf{C}^{+}_{i}$ is the control signal generated by the *ControlNet*. We represent the current state of motion for the human model: $\mathbf{J}_i = (\boldsymbol{j}^{p}_{i}, \boldsymbol{j}^{v}_{i}, \boldsymbol{j}^{r}_{i})$ contains root relative position $\boldsymbol{j}^{p}_{i} \in R^{j\times 3}$, rotation $\boldsymbol{j}^{v}_{i} \in R^{j\times 6}$ and velocity $\boldsymbol{j}^{r}_{i} \in R^{j\times 3}$ of each joint at frame i. We use $j = 22$ joints for our human model. $\mathbf{T}_i = (\boldsymbol{t}^{p}_{i}, \boldsymbol{t}^{d}_{i}, \boldsymbol{t}^{a}_{i})$ contains the root positions $\boldsymbol{t}^{p}_{i} \in R^{\mathcal{T}\times 3}$ and rotation $\boldsymbol{t}^{d}_{i} \in R^{\mathcal{T}\times 6}$ for $\tau = 13$ frames sampled uniformly between the [-1, 1] second window centered at frame i. $\boldsymbol{t}^{a}_{i} \in R^{\mathcal{T}\times 3}$ are the soft labels which describe current action over ours three action classes, namely, idle, walk, and sit. Inspired by Starke et al., [47], we also use intermediate goals $\mathbf{G}_i = (\boldsymbol{g}^{p}_{i}, \boldsymbol{g}^{d}_{i}, \boldsymbol{g}^{a}_{i})$, where $\boldsymbol{g}^{p}_{i} \in R^{\mathcal{T}\times 3}$, $\boldsymbol{g}^{d}_{i} \in R^{\mathcal{T}\times 6}$ are the goal positions and orientations at frame i. $\boldsymbol{g}^{a}_{i} \in R^{\mathcal{T}\times 3}$ are the one-hot labels describing the intended goal action.

To accurately capture the spatial relation between the person and the chair, we voxelize the chair into an $8\times 8\times 8$ grid and store at each voxel its occupancy ($\mathbb{R}$) and the relative vector between the root joint of the person and the voxel ($\mathbb{R}^3$). This allows us to reason about the distance between the person and different parts of the chair. We flatten this grid to obtain our chair encoding $\mathbf{I}_i \in \mathbb{R}^{2048}$ at time-step i.

In order to explicitly reason about the collisions of the person with the chair, we voxelize the region around the person into a cylindrical ego-centric grid and store the occupancies corresponding to the chair (if it is inside the grid). We flatten the occupancy feature to obtain $\mathbf{E}_i \in \mathbb{R}^{1408}$. It is important to note that although $\mathbf{I}_i$ and $\mathbf{E}_i$ are scene encodings that serve different purposes. $\mathbf{I}_i$ is chair-centric and entails information about how far is the person from the chair and the geometry of the chair, while $\mathbf{E}_i$ is ego-centric and detects collisions in the surrounding of the human model. In addition, we also introduce an auxiliary variable $\mathbf{\Phi} \in [0, 1]$ as in [41,27], which encodes the global phase of the motion. When approaching the goal, the represents the timing within a walking cycle, for sitting the phase equals 0 when the person is still standing and reaches 1 when the person has sat.

The components of the output of the network differs from the input to a small extend by additionally predicting $\tilde{\boldsymbol{j}}^{p}_{i+1}$ are the joint positions relative to future root 1 second ahead. To ensure the human model can reach the chair, we introduce the goal-relative root trajectory $\tilde{\mathbf{T}}_{i+1} = \{\tilde{\boldsymbol{t}}^{p}_{i+1}, \tilde{\boldsymbol{t}}^{d}_{i+1}\}$ which include the root positions and forward directions relative to the chair of frame $i + 1$. The rest of the components remain consistent with the input include the the future pose $\mathbf{J}_{i+1}$, future root trajectory $\mathbf{T}^{+}_{i+1}$, the future intermediate goals $\mathbf{G}^{+}_{i+1}$, and the future global phase $\mathbf{\Phi}_{i+1}$. The PoseNet $f^{\mathrm{PN}}(\cdot)$ adopts a mixture-of-experts [27,22,47,48] and is trained to minimize the standard MSE loss.

4.4 Contact Generation with ContactNet

From the user's perspective, it is useful to automatically generate plausible contact points on any given chairs. To this reason, we propose *ContactNet*. The

network adopts a conditional variational auto-encoder[46] architecture (cVAE) which encodes the chair geometry $\mathbf{I}$ introduced in Section 4.3 and the contact positions $\boldsymbol{c} \in R^{2\times 3}$ to a latent vector $\boldsymbol{z}$. The decoder of the network then reconstructs the hand contacts $\hat{\boldsymbol{c}} \in R^{2\times 3}$. Note, the position of each voxel in the scene representation $\mathbf{I}$ in this case is computed relative to the center of the chair instead of the character's root. During training, the network is trained to minimize the following loss,

$$L_{\text{contact}} = \|\hat{\boldsymbol{c}} - \boldsymbol{c}\|_2^2 + \beta KL(q(\boldsymbol{z}|\boldsymbol{c}, \mathbf{I})\|p(\boldsymbol{z})), \tag{4}$$

where KL denotes the Kullback-Leibler divergence. During inference, given the scene representation $\mathbf{I}$ of a novel chair, we sample the latent vector $\boldsymbol{z}$ from the uniform Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and use the decoder to generate plausible hand contacts $\boldsymbol{c} \in \mathbb{R}^{2\times 3}$.

4.5 Decomposition of *Approaching* and *Sitting* Motion

Detailed hand motion planning is only required when the human model is close enough to the chair right before sitting as sitting requires synthesizing more precise full-body motion, especially for the hands, such that the person makes the desired contacts and sits on the chair. For this reason, we decompose our synthesis into approaching and sitting by only activating the ControlNet during the sitting. When the ControlNet is deactivated the control signal or when a "no contact" signal is present the control signal for the corresponding hand is zeroed.

5 Evaluation

Studying contact-conditioned interaction with chairs requires accurately labelled contacts and a diverse range of styled interactions. The COUCH dataset is captured to meet such needs. We evaluate our contact constrained motion synthesis method on the COUCH dataset qualitatively and quantitatively. Our method is the first approach that allows the user to explicitly define how the person should contact the chair and we generate natural and diverse motions satisfying these contacts. As such we evaluate our method on three axis, (i) accuracy in reaching the contacts, (ii) diversity and (iii) naturalness of the synthesised motion. For qualitative results, we highly encourage the readers to see our supplementary video. It can be seen that our method can generate diverse and natural motions while reaching the user-specified contacts. We quantitatively evaluate the accuracy of contacts and motion diversity on a total of 120 testing sequences on six subject-specific models trained on corresponding subsets of our COUCH dataset. Note that we evaluate raw synthesized motion without post-processing.

Fig. 4: We demonstrate qualitatively and quantitatively (Tab. 2) that motion generated by our approach satisfies the contacts much better than the baselines, NSM+Control [47] and SAMP+Control [22].

5.1 Baselines

To our best knowledge, the most related work to ours are the Neural State Machine (NSM) [47] and the SAMP [22] since they both synthesize human-scene interactions. However, neither of the methods allows the use of fine-grained control over how the interaction should take place. We adapt these baselines for our task by additionally conditioning on the contact positions and refer to these new baselines as NSM+Control and SAMP+Control. Quantitative results are reported for both the original baselines and their adapted version. For each of the methods, we train subject-specific models with the corresponding subset of our COUCH dataset using the code provided by the authors. Our experiments, detailed below, show that naively providing contacts as input to existing motion synthesis approaches does not ensure that the generated motion satisfies the contacts. Our method, on the other hand, does not suffer from this limitation.

5.2 Evaluation on Control

In order to evaluate how well our synthesised motion meets the given contacts, we report the *average contact error (ACE)* as the mean squared error between the predicted hand contact and the corresponding given contact. We use the closest position of the predicted hand motion to the given contact as our hand contact.
Since ACE might be susceptible to outliers and inspired by the literature on object detection [35,43,42,11], we also report *average contact precision* (AP@k), where we consider a contact as correctly predicted if it is closer than k cm.

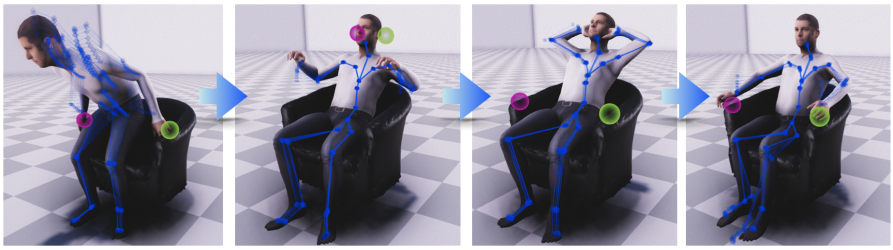

Fig. 5: COUCH can also be extended by specifying a series of contacts for to automatically synthesize more complex interactions. The past poses are indicated by blue skeletons.

We compare our method with NSM+Control and SAMP+Control in Table 2. It can be observed that COUCH outperforms prior methods by a significant margin. Prior methods are trained to condition on the contact positions, however it is found (Figure 4) to be not sufficient as the contact input can be easily ignored during auto-regressive prediction. As a result the contact constraints are often not met. This highlights the importance of motion planning in form of trajectory predictors in order to reach the desired contacts. Our ControlNet provides valuable information on how to synthesise motion such that the given contacts are satisfied. Our motion prediction network PoseNet uses these control signals to generate contact constrained motions.

Table 2: Evaluation on degree of control. COUCH is shown to be more controllable compared to the baseline methods. The distance from given contact points and the joint position are measured. The success rate of control is also reported.

Method	Distance to Contact$^{\downarrow}$	AP@ 3 cm$^{\uparrow}$	AP@ 5 cm$^{\uparrow}$	AP@ 7 cm$^{\uparrow}$	AP@ 9 cm$^{\uparrow}$
NSM [47]	10.69	15.52	38.20	46.05	56.61
SAMP [22]	11.96	6.54	14.57	20.94	50.83
NSM+Control [47]	10.52	17.46	35.7	48.4	57.93
SAMP+Control [22]	12.09	7.20	15.2	23.2	48.80
Ours	**4.73**	**47.97**	**78.86**	**87.8**	**91.87**

5.3 Evaluation on Motion Diversity

Diversity is an essential element for our motion synthesis, since a chair can be approached and interacted with in different ways. To quantify diversity, we evaluate using the Average Pairwise Distance (APD) [55,58,22] on the synthesized pose features of the virtual human $\mathbf{J}_i = (\boldsymbol{j}_i^p, \boldsymbol{j}_i^v, \boldsymbol{j}_i^r)$. defined as:

$$APD = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i}^{N} D(\mathbf{J}'_i, \mathbf{J}'_j), \tag{5}$$

where N is the total number of frames in all the testing sequences. Note that for evaluation, the virtual human is initialized at different starting points and is instructed to approach and sit on randomly selected chairs with randomly sampled contact points from the dataset, and motion is synthesized for 16 seconds for each sequence. We compare the diversity of synthesized motion in Table 3 and it can be seen that using explicit contacts allows our method to generate more varied motion.

Table 3: Evaluation on the diversity of the synthesized motion. APD is measured for segmented motion of approaching and sitting. Our approach attains the best score compared to the baselines.

Method	Approach	Sit
NSM [47]	5.15	5.76
SAMP [22]	5.34	5.81
NSM+Control [47]	5.07	5.80
SAMP+Control [22]	5.21	5.88
Ours	**5.55**	**6.02**
Ground Truth	5.69	6.30

5.4 Controlling with a series of Contacts.

A useful application of our approach is to automatically generate a motion sequence with a series of desired contacts in the context of animation, character control, when executing a set of complex actions. For instance, the person can be instructed to first sit with their hands on the armrest, then lift the arms to support the head before bringing the hands back to the armrest, see Figure 5) and the supplementary video. Our approach can be adapted for this task by iteratively providing the new goal locations for the hands as input after the present locations are reached.

5.5 Contact Prediction on Novel Shapes

Apart from user-specified contacts, we can additionally generate the contacts on the surface of a given chair using our proposed ContactNet. This allows us to generate fully automatic and diverse motions for sitting. To measure the diversity of the generated contacts from ContactNet, we compute the Average Pairwise Distance (APD) among the generated hand contact positions c_j with unseen chair shapes. A total number of 200 unseen chairs are chosen, and each 10 contact positions are predicted for both hands.

$$APD = \frac{1}{2LN(N-1)} \sum_{k=1}^{2} \sum_{l=1}^{L} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \|X_i' - X_j'\|_2^2 \tag{6}$$

$L = 200$ is the number of objects and $N = 10$ is the number of contacts generated per object. The APD on contact positions is **11.82** cm which is comparable to

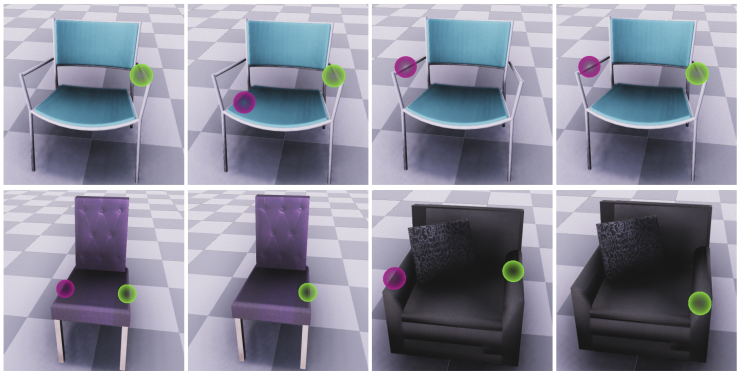

Fig. 6: *ContactNet* enables sampling of diverse contact positions across various chair shapes. These contacts can be used by our ControlNet and PoseNet to generate fully automatic and diverse motions.

the ground truth dataset which has an APD of **14.07** cm. As shown qualitatively in Figure 6, we can generate diverse and plausible contact positions on chairs, which can generalize to unseen shapes.

6 Conclusion

We propose COUCH, the first method for synthesising controllable contact-driven human-chair interactions. Given initial conditions and the contacts on the chair, our model plans the motion of the hands, which drives the full body poses to satisfy contacts. In addition to the model, we contribute the COUCHdataset for human chair interactions which includes a wide variety of sitting motions approaching and contacting the chair in different ways. It consists of 3 hours of motion capture with 6 subjects interacting with registered 3D chair models, captured in high quality with IMUs and Kinects. Experiments demonstrate that our method consistently outperforms the SoTA by improving the average contact accuracy by ∼55% to better satisfy contact constraints. In addition to better control, it can been seen in the supplementary video that our approach generates more natural motion compared to the baseline methods. In the future, we want to extend our dataset to new activities and train a multi-activity contact driven model. In the supplementary, we discuss further future directions in this new problem of fine-grained controlled motion synthesis. Our dataset and code will be released to foster further work in this new research direction.

Acknowledgement

We would like to thank Xianghui Xie for helping with data processing, and we are very grateful for all the participants who took part in the data capture. This work is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans). Gerard Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. The project was made possible by funding from the Carl Zeiss Foundation.

References

1. http://virtualhumans.mpi-inf.mpg.de/couch/
2. https://www.treedys.com/.
3. Agisoft metashape. https://www.agisoft.com/
4. Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors. https://www.xsens.com/, accessed: 2010-09-30
5. Aksan, E., Kaufmann, M., Hilliges, O.: Structured prediction helps 3d human motion modelling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
6. Aliakbarian, S., Saleh, F.S., Salzmann, M., Petersson, L., Gould, S.: A stochastic conditioning scheme for diverse human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
7. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single RGB camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (jun 2019)
8. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: European Conference on Computer Vision (ECCV). Springer (August 2020)
9. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: ContactDB: Analyzing and predicting grasp contact via thermal imaging (2019), cVPR
10. Cao, Z., Gao, H., Mangalam, K., Cai, Q., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: ECCV (2020)
11. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
12. Chang, A.X., Funkhouser, T.A., Guibas, L.J., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository. CoRR **abs/1512.03012** (2015)
13. Chao, Y., Yang, J., Chen, W., Deng, J.: Learning to sit: Synthesizing human-chair interactions via hierarchical control. CoRR **abs/1908.07423** (2019)
14. Corona, E., Pumarola, A., Alenyà, G., Moreno-Noguer, F.: Context-aware human motion prediction. CoRR **abs/1904.03419** (2019)
15. Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F., Rogez, G.: Ganhand: Predicting human grasp affordances in multi-object scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
16. Cui, Q., Sun, H., Yang, F.: Learning dynamic relationships for 3d human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
17. Eigen, D., Ranzato, M., Sutskever, I.: Learning factored representations in a deep mixture of experts. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings (2014)
18. Ghosh, P., Song, J., Aksan, E., Hilliges, O.: Learning human motion models for long-term predictions. s International Conference on 3D Vision 3DV (2017)
19. Gui, L.Y., Wang, Y.X., Liang, X., Moura, J.M.F.: Adversarial geometry-aware human motion prediction. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)

20. Gui, L.Y., Wang, Y.X., Ramanan, D., Moura, J.M.F.: Few-shot human motion prediction via meta-learning. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
21. Habibie, I., Holden, D., Schwarz, J., Yearsley, J., Komura, T.: A recurrent variational autoencoder for human motion synthesis. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 119.1–119.12. BMVA Press (September 2017)
22. Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.: Stochastic scene-aware motion prediction. In: Proceedings of the International Conference on Computer Vision 2021 (2021)
23. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: International Conference on Computer Vision. pp. 2282–2292 (Oct 2019)
24. Henter, G.E., Alexanderson, S., Beskow, J.: Moglow: probabilistic and controllable motion synthesis using normalising flows. ACM Trans. Graph. **39**(6), 236:1–236:14 (2020)
25. Hernandez, A., Gall, J., Moreno-Noguer, F.: Human motion prediction via spatio-temporal inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
26. Holden, D., Kanoun, O., Perepichka, M., Popa, T.: Learned motion matching. ACM Trans. Graph. **39**(4), 53 (2020)
27. Holden, D., Komura, T., Saito, J.: Phase-functioned neural networks for character control. ACM Trans. Graph. **36**(4), 42:1–42:13 (2017)
28. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Trans. Graph. (Jul 2016)
29. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (jul 2014)
30. Juliani, A., Berges, V., Vckay, E., Gao, Y., Henry, H., Mattar, M., Lange, D.: Unity: A general platform for intelligent agents. CoRR **abs/1809.02627** (2018)
31. Karunratanakul, K., Yang, J., Zhang, Y., Black, M., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. In: International Conference on 3D Vision (3DV) (2020)
32. Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., Tian, Q.: Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
33. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++ (2021)
34. Li, X., Liu, S., Kim, K., Wang, X., Yang, M.H., Kautz, J.: Putting humans in a scene: Learning affordance in 3d indoor environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
35. Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: CVPR. pp. 936–944. IEEE Computer Society (2017)
36. Ling, H.Y., Zinno, F., Cheng, G., Van De Panne, M.: Character controllers using motion vaes. ACM Trans. Graph. **39**(4) (2020)

37. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (2015)
38. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: archive of motion capture as surface shapes. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019
39. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
40. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017
41. Pavllo, D., Grangier, D., Auli, M.: Quaternet: A quaternion-based recurrent model for human motion. In: British Machine Vision Conference (BMVC) (2018)
42. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28 (2015)
43. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. CoRR **abs/1506.01497** (2015)
44. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In: IEEE International Conference on Computer Vision Workshops (2021)
45. Sofiiuk, K., Petrov, I., Barinova, O., Konushin, A.: f-brs: Rethinking backpropagating refinement for interactive segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8623–8632 (2020)
46. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. pp. 3483–3491 (2015)
47. Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Trans. Graph. **38**(6), 209:1–209:14 (2019)
48. Starke, S., Zhao, Y., Komura, T., Zaman, K.A.: Local motion phases for learning multi-contact character movements. ACM Trans. Graph.
49. Starke, S., Zhao, Y., Zinno, F., Komura, T.: Neural animation layering for synthesizing martial arts movements. ACM Trans. Graph.
50. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV) (2020)
51. Wang, H., Feng, J.: VRED: A position-velocity recurrent encoder-decoder for human motion prediction. CoRR **abs/1906.06514** (2019)
52. Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 9401–9411. Computer Vision Foundation / IEEE (2021)
53. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)

54. Xu, J., Xu, H., Ni, B., Yang, X., Wang, X., Darrell, T.: Hierarchical style-based networks for motion synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI
55. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
56. Zhang, H., Starke, S., Komura, T., Saito, J.: Mode-adaptive neural networks for quadruped motion control. ACM Trans. Graph. **37**(4), 145:1–145:11 (2018)
57. Zhang, S., Zhang, Y., Ma, Q., Black, M.J., Tang, S.: PLACE: Proximity learning of articulation and contact in 3D environments. In: International Conference on 3D Vision (3DV) (Nov 2020)
58. Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3d people in scenes without people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

APPENDIX

In this appendix, we provide additional information about the dataset, implementation details, post-processing techniques. We also discuss on the current limitation as future research perspectives.

1 Dataset

1.1 Motion Data

Table 4 shows a break down of our dataset in terms of different types of interactions. Our dataset consists of 3 hours of MoCap with over 500 motion sequences.

Table 4: Distribution of the COUCH dataset with different types of interaction.

Interaction Type	Minutes	%
Right Hand	36.3	17.3
Left Hand	29.4	14.0
Both Hand	60.5	28.9
No Contact	36.5	17.4
Free Interaction	31.9	15.2
Locomotion	15.1	7.2

1.2 Data Processing

SMPL Fitting. We segment the human in captured RGB images by running Detectron V2 [53] followed by manual correction with [45] on the segmentation masks. These masks are then used to segment multi-view depth maps and lift human point clouds from 2D to 3D. We use FrankMocap [44] to initialize the SMPL pose from the images and then apply instance specific optimization [7] to fit the SMPL model to the segmented human point cloud. For more accurate fitting, we additionally obtain the SMPL shape parameters of each subject from 3D scans using [8].

Synchronization with the IMUs. The fitted SMPL model provides us with accurate contacts with the scene, however, the fitted motion sequence is prone to occlusion and drastic body movements, as a result, the fitted motion can be jittery at times. On the other hand, the pose captured with the IMUs is smooth over time, but it might not accurately capture the contacts. To this reason, we synchronize the Kinect captured data with the body sensors by incorporating the SMPL fitted poses into the IMU pose sequences. After synchronization, we optimize the joint rotations $\boldsymbol{j}_i^r$ to achieve temporal smoothness via the objective

$$L_{\text{temp}}(\boldsymbol{j}_i^r) = \sum_{i=1}^{T-1} \|\boldsymbol{j}_{i+1}^r - \boldsymbol{j}_i^r\|^2 + \sum_{i=1}^{T-1} \|\ddot{\boldsymbol{j}}_i\| \tag{7}$$

where $\ddot{\boldsymbol{j}}_i$ represents the acceleration of the body joints in frame i approximated by central difference.

We additionally use the binary contact labels of the toes and the heels detected by the IMU sensors to remove foot-sliding on the motion data. To remove the foot-sliding, we compute the average joint positions over the duration of the contacts grouped by the positive contact labels. This computation is performed for all four foot joints. This forms a sequence of target joint positions of the feet $\tilde{\boldsymbol{f}}_i \in R^{4\times 3}$. We then optimize the objective function

$$L_{\text{slide}}(\boldsymbol{j}_i^r) = \sum_{i=1}^{T} \|\tilde{\boldsymbol{f}}_i - \boldsymbol{f}_i\|^2, \tag{8}$$

where $\boldsymbol{f}_i$ represents the foot joint positions at frame i. The resulting motion sequence is temporally smooth and has accurate contacts registered with the chair models.

Object Processing. To obtain object segmentation, we pre-scan objects using a 3D scanner [2,3]. We then use multi-view object keypoints, marked by manual annotators on the images, to fit the pre-scanned chair meshes to the given frame. The segmentation masks are then obtained by projecting fitted object meshes to the images. Since the chairs remain static during the capture, we average over the 6D pose of the fitted chair model during each capture session to obtain the final transformation of the chair.

2 Training Details

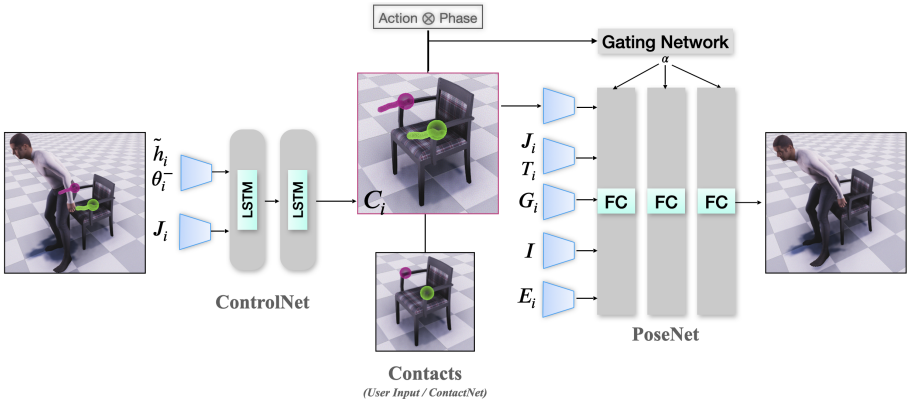

Fig. 7: Our method that combines the ControlNet and the PoseNet.

2.1 ControlNet

As shown in Figure 7, the contact network is a two-layer LSTM architecture. Each layer has a hidden dimension of 512. The pose and the control signals (hand trajectories, and the local phases) are each encoded through a two-layer fully connected network with of shape {128, 128} before passing through the LSTM. We apply scheduled sampling on hand trajectories for better model performance. For the local phases, we always use the ground truth. Each of our training samples is in a sequence of 60 frames. The ControlNet is trained for 150 epochs with an Adam optimizer. The initial learning rate is 1e-3 and a cosine learning rate scheduler was used to decay the learning rate gradually to 5e-6. The full training of a subject-specific model takes approximately 1 hour on an NVIDIA V100 GPU.

2.2 PoseNet

Table 5: Details on different encoder networks of the PoseNet.

Networks	Architecture
Encoder for **C**	{128,128,128}
Encoder for {**J**, **T**}	{512, 512, 512}
Encoder for **G**	{128,128,128}
Encoder for **I**	{512, 512, 512}
Encoder for **E**	{256,256,256}

The PoseNet adopts the mixture-of-expert structure [17]. It consists of different feature encoders of structures shown in Table 5. The gating network and the prediction networks are both three-layer fully-connected networks, with hidden dimensions of 128 and 512 respectively. The number of experts is set to 10. The PoseNet is trained for 150 epochs with an Adam optimizer. The initial learning rate is 1e-4 and a cosine learning rate scheduler was used to decay the learning rate gradually to 5e-6. The full training of a subject-specific model takes approximately 6 hours on an NVIDIA V100 GPU.

2.3 ContactNet

The ContactNet encodes the scene **I** through a three-layer fully connected network of shape {512, 512, 64}. The latent vector $\boldsymbol{z}$ of the VAE is of size 6. The weight of the Kullback-Leibler divergence β is 0.1. We use the Adam optimizer with a learning rate of 1e-3 and train ContactNet for 150 epochs. The full training of a subject-specific model takes approximately 10 minutes on an NVIDIA V100 GPU.

3 Contact Projection and Trajectory Fitting

To ensure the ContactNet predicts contacts that land exactly on the surface of the object, we perform a post-processing step, when the distance of the network predicted contact to the surface is less than a set threshold of 10 cm, we simply project the contact onto the nearest point on the chair surface. When the distance is greater than 10 cm, we simply neglect the predicted contact. The ControlNet predicts the future hand trajectories, and it would be possible to fit the predicted pose to the predicted hand position from the hand trajectories at each frame to further improve the satisfaction of the contact constraints. Note, in the evaluation of the main paper we do not apply such fitting technique.

4 Limitations and Future Direction

We observe the synthesized motion can slightly intersect with the chair. A solution to this problem would be to apply a post-processing step to avoid such collision. In order to generalize to more different chair shapes, it would be useful to investigate better ways of encoding the scene geometry while trying to avoid over-fitting.

Different shaped person can intersect with the same object very differently even when performing the same motion. The COUCH dataset captures human interaction with different body shapes. With the dataset, it is possible to study how to build subject-variant motion synthesis model and how to effectively condition on the body shapes. These are challenges in motion synthesis that have not been tackled.

Our work on controllable human-chair interaction. It would be useful to extend the scope of interacted objects, especially considering the cases when the objects are non-static, when performing motions such as lifting a box, or opening a door. Another possible direction would be to further apply contact-based control in these interactions.