

# Generating Continual Human Motion in Diverse 3D Scenes

Aymen Mir<sup>1,2</sup>    Xavier Puig<sup>3</sup>    Angjoo Kanazawa<sup>4</sup>    Gerard Pons-Moll<sup>1,2</sup>

<sup>1</sup> Tübingen AI Center, University of Tübingen, Germany

<sup>2</sup> Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

<sup>3</sup> Meta AI Research,

<sup>4</sup> University of California, Berkeley

{aymen.mir,gerard.pons-moll}@uni-tuebingen.de, xavierpuig@meta.com, kanazawa@eecs.berkeley.edu



Figure 1. Our method synthesizes diverse animator guided human motion such as sitting and grabbing in diverse 3D scenes. We urge readers to watch the supplementary video for more results.

## Abstract

We introduce a method to synthesize animator guided human motion across 3D scenes. Given a set of sparse (3 or 4) joint locations (such as the location of a person’s hand and two feet) and a seed motion sequence in a 3D scene, our method generates a plausible motion sequence starting from the seed motion while satisfying the constraints imposed by the provided keypoints. We decompose the continual motion synthesis problem into walking along paths and transitioning in and out of the actions specified by the keypoints, which enables long generation of motions that satisfy scene constraints without explicitly incorporating scene information. Our method is trained only using scene ag-

nostic mocap data. As a result, our approach is deployable across 3D scenes with various geometries. For achieving plausible continual motion synthesis without drift, our key contribution is to generate motion in a goal-centric canonical coordinate frame where the next immediate target is situated at the origin. Our model can generate long sequences of diverse actions such as grabbing, sitting and leaning chained together in arbitrary order, demonstrated on scenes of varying geometry: HPS, Replica, Matterport, ScanNet and scenes represented using NeRFs. Several experiments demonstrate that our method outperforms existing methods that navigate paths in 3D scenes.

## 1. Introduction

Our goal is to generate animator guided rich long-term human behavior in arbitrary 3D scenes, including a variety of actions and transitions between them. Such a system should allow for goal-directed generation of humans moving about from one place to another, for example, walking towards the couch to sit on it, and then standing up and approaching the shelf to grab something from it, as illustrated in Figure 1. It should allow users to specify with minimal interaction what kind of actions to perform, while keeping the realism and expressivity required for applications such as synthetic data generation, robotics, VR/AR, gaming, etc.

While the community has seen promising progress in animator guided motion synthesis in 3D scenes, most works are restricted to a single action and do not handle transitions [73, 69, 62], preventing them from producing long range diverse motion. They are also not deployable in a wide variety of real scenes [58, 65, 66, 26]. The reason for this is that they synthesize motion by conditioning on scene geometry and require training on a dataset featuring 3D humans interacting in 3D scenes and objects [27, 26, 73]. Generalizing these methods to arbitrary 3D scenes would require collecting motion data registered to a myriad of possible 3D scenes and objects, which is not scalable.

In contrast, humans can navigate cluttered scenes, pick objects from a shelf they have never seen before, and sit on novel furniture and surfaces. Most of the clutter in the scene is often ignored, and what matters most are not the exact details of the object / scene geometry but whether they afford each action. Our hypothesis is that motion, to a large extent, is driven to avoid obstacles and focused on reaching the next immediate goal or target contacts in the environments. Thus, it should be possible to generate human motion without accounting for all the details in the 3D scene.

Based on this insight, we propose a novel framework for animator-guided motion synthesis in 3D scenes without relying on scene-registered motion data. As such, our method can be trained on regular mocap data, which is relatively easily captured and abundantly available [46]. Since our method does not explicitly condition on the geometry of the scene, it can be deployed across 3D scenes with varied geometry.

Our method relies on two key observations: first, we can represent actions in a 3D scene as a set of sparse desired target contacts (we use 3 or 4 contacts such as the location of the two feet and a hand or the location of two feet and the root) to be reached, which we refer to as *action keypoints*. These keypoints can be provided by an animator using an interface or generated by automated heuristics, allowing animators to trade off the speed and control over the generation motion. An interesting finding in this paper is that *action keypoints* are a powerful abstraction of several actions in 3D scenes, and can be used to execute instruc-

tions such as “sit there” or “grab at this height”. Second, avoiding obstacles in 3D scenes can be achieved by path following. The challenge is to follow arbitrarily long paths, smoothly making the human transition into and out of the action, and then walk towards the next target. For this, we break down motion into three pieces: walking, transition into and out of an action. For path following and transitions, we introduce the idea of training a motion synthesis model entirely with *scene-agnostic motion data* to reach the origin of a *canonical coordinate frame*. For navigating paths, this model is sampled iteratively to converge at the origin of the *canonical coordinate frame* defined using waypoints and tangents on the path. For transitions in and out of actions, motion is synthesized by placing target poses at the origin of the canonical coordinate frame. By iteratively synthesizing motion in the *canonical coordinate frame*, our method allows for long range motion synthesis that transitions between walks and various actions in a 3D scene.

Unlike existing methods for motion synthesis [26, 58], our method allows for synthesizing motion without requiring any manual phase or action annotation.

For the first time, we demonstrate long-range human motion synthesis on a wide range of scene datasets: Replica [61], Matterport [8], HPS [22], Scannet [10] and a NeRF scene. Furthermore, we show that our model can perform actions at different places, such as grabbing from any shelf, table or cabinet at any height or sitting on any surface that affords sitting. We will make our code and models publicly available which can be used by animators to synthesize goal directed human motion across 3D scenes.

To summarize, our contributions are as follows:

- We present a method that departs from existing methods for motion synthesis in 3D scenes by only using regular motion capture data and that is deployable across varied 3D scenes.
- We introduce a novel idea of iteratively converging motion at the origin of a canonical coordinate frame, which allows to synthesize long-range motion in 3D scenes.

## 2. Related Work

**Human Motion Prediction without the 3D scene.** Predicting the dynamics of human motion is a long studied problem in computer vision and graphics. Classic works explored using Hidden Markov Chains [5] and Gaussian Processes [64], physics based models [44] for predicting future motion. Recently, recurrent neural networks [19, 31] have been used for motion prediction [17, 48, 3] also in combination with Graph Neural Networks [36, 47, 40, 11], and variational Auto-encoders [35] to add diversity [23, 74]. Yuan et al. [71]. An intrinsic problem of recurrent methods is that they drift over time [1].

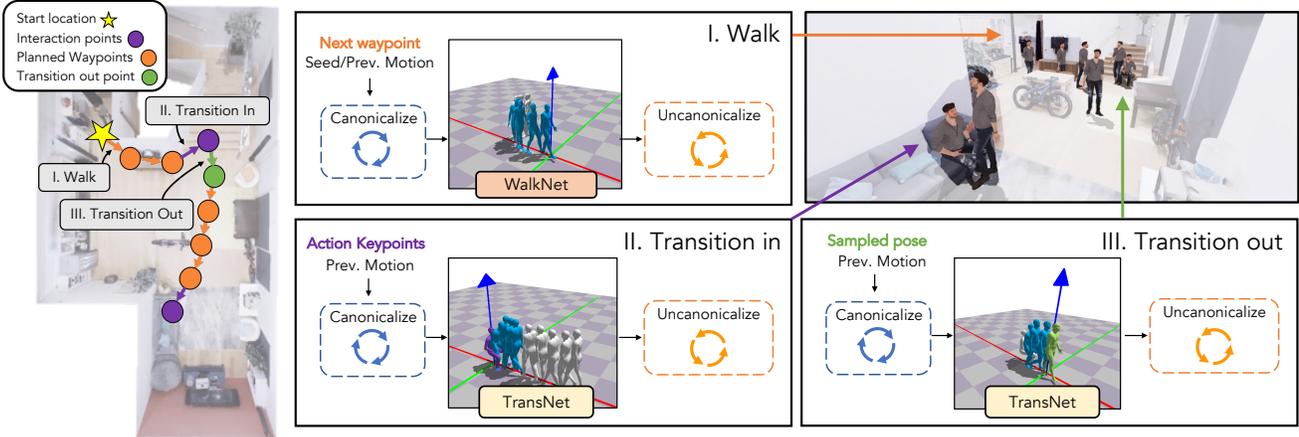


Figure 2. Overview of our method. We generate human motion satisfying keypoint constraints by dividing it into 3 stages: a *Walk Motion*, which animates the human as it walks between keypoints, a *Transition-In*, which blends the walking motion with the pose specified by the keypoints and a *Transition-Out*, which animates the human back to the walking pose. We use an autoregressive transformer, *WalkNet*, to synthesize the walking motion, and a masked-autoencoder transformer to generate the blending motion. By moving the motion into a Goal-Centric Canonical Coordinate Frame our method can generalize to a wide set of 3D scenes.

More recent approaches employ transformers to generate unconditional or text and music conditioned motion sequences [1, 41, 39, 51, 52]. We also build on transformer architectures but aim to generate motion in real 3D scenes.

Motion Inbetweening [14, 25, 49, 70, 2, 34] is another classic paradigm for motion synthesis where the task is to fill in frames between animator provided keyframes.

Our approach builds on recent progress in transformer architectures [41], and classical ideas such as motion inbetweening, combined with the novel idea of a canonical coordinate frame and action keypoint representation in order to generate motion in 3D scenes.

**Character Control in Video Games.** Motion matching [54], its learnt-variant [9, 32] and motion graphs [38, 15, 37, 56, 55] are classical methods often employed in the industry for generating kinematic motion sequences, controlled by environment and user specified constraints. Similar to our goal, some works [53, 7] use a combination of these approaches and IK to generate human behaviors in synthetic scenes. However, these approaches require significant human effort to author realistic animations, and IK approaches easily produce non-realistic animations.

Deep learning variants such as Holden *et al.* [33] introduce phase-conditioning in a RNN to model the periodic nature of walking motion. In several works by Starke *et al.* [58, 60, 59] the idea of local phases is extended to synthesize scene aware motion, basketball motion and martial arts motion. All these methods generate convincing motion but phases are non-intuitive for non-periodic motion and often require manual labelling.

**Static Human Pose Conditioned on Scenes.** The relationship between humans, scenes, and objects is another recurrent subject of study in computer vision and graphics. Classical works include methods for 3D object detection [20, 21] and affordance prediction using human poses [12, 18, 16].

Several recent works, generate plausible static poses conditioned on the a 3D scene [42, 74, 68, 72, 28, 76] using recently captured human interaction datasets [27, 22, 57, 4, 63, 6]. Instead of static poses, we generate *motion* coherent with the scene which is significantly harder.

**Scene Aware Motion Synthesis.** Some works leverage reinforcement learning to synthesize navigation in *synthetic* 3D scenes [43, 75]. Other works focus on a single action, such as grabbing [62, 69] but do not demonstrate transitions to new motions. These methods are not demonstrated in real 3D scenes with multiple objects and clutter. Recent real interaction datasets [27, 22, 57, 4, 63, 6] have powered methods to synthesize 3D scene aware motion [66, 65, 6, 67]. These datasets are crucial to drive progress, but do not capture the richness and variety of real world scenes. Hence, these methods are often demonstrated only on small scenes from PROX [27] and Matterport [8].

We draw inspiration from Hassan *et al.* [26] which combine path planning with neural motion synthesis, and from Zhang *et al.* [73] which synthesize contact controlled human chair interaction. These methods require the geometry of the isolated interacting object as input, which make them hard to generalize to real 3D scenes. Unlike these methods, we demonstrate *long chained sequences of actions* in *generic real 3D scenes*, which is enabled with our origin

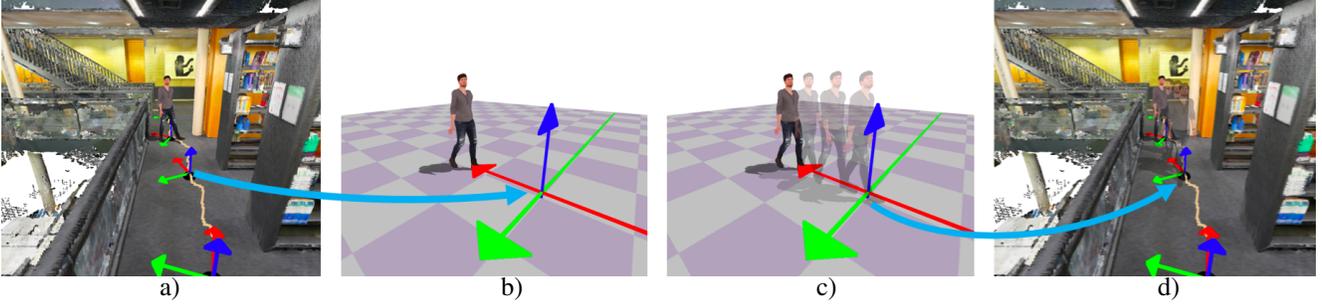


Figure 3. a) Using keypoints and tangents along a path, we move motion from the scene coordinate frame into b) the goal-centric canonical coordinate frame, where c) *WalkNet* synthesizes motion that converges at the origin of the coordinate frame. d) Once the synthesized motion reaches the origin, we move it back to the scene coordinate frame.

canonicalization and action keypoints.

### 3. Method

Our method takes as input a seed motion sequence and a list of action keypoints  $\{\mathbf{a}_1, \dots, \mathbf{a}_n\}$  specifying interactions at different locations in the scene. These keypoints can be specified by users or generated using language commands and scene segmentations (Sec. 3.2). Our goal is to synthesize motion that starts at the seed motion and transitions in and out each of the action keypoints in the input list.

The first step is to optimize for a pose that fits the action keypoints at target locations using Inverse Kinematics and a pose-prior (3.3). These poses along with the starting seed motion act as anchors to guide the motion synthesis process.

Using scene-agnostic motion capture data placed in a goal-centric canonical coordinate frame (Sec. 3.4), we train *Walknet* (Sec. 3.5) to synthesize walking motion that converges at the origin of a canonical coordinate frame, and *TransNet* (Sec 3.6) that synthesizes motion inbetween a seed motion sequence and a target pose also at the origin. At test time (see Fig. 2), *WalkNet* is used to reach canonicalized intermediate goals along a path computed with a path planning algorithm, thus creating long motion by successively reaching the origin. Once the walking motion reaches the vicinity of an anchor pose, *TransNet* synthesizes transition from walking motion to the anchor pose and vice versa. This allows to synthesize motion in 3D scenes without the need for motion data coupled with 3D scenes. Our framework is general and highly modular, which allows it to be updated with novel methods for motion synthesis.

#### 3.1. SMPL Body Model

We use the SMPL body model [45] to represent the human subject. SMPL is a differentiable function  $M(\phi, \theta, \mathbf{t}, \beta)$  that maps global body orientation  $\phi$ , pose  $\theta$ , translation  $\mathbf{t}$  and shape  $\beta$  parameters to the vertices of a human mesh along with the 3D joint locations of the SMPL skeleton. We assume that  $\beta$  remains static through-

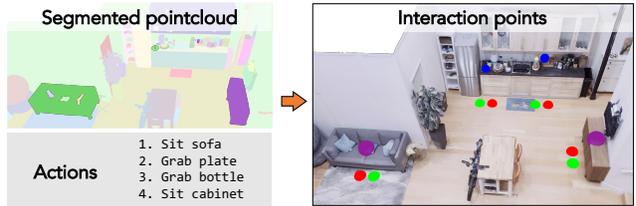


Figure 4. Using language instruction and semantic segmentation, keypoints can be automatically placed in a 3D scene.

out our method. We denote motion sequences as an ordered list of SMPL parameter tuples. For example  $\mathcal{C} = [(\mathbf{r}, \phi, \theta)_j]_{j=1:D}$  denotes a motion sequence of  $D$  frames.

#### 3.2. Generating Keypoints in a Scene

Keypoints can be efficiently collected using a 3D user interface, as described in the supp. mat, or keypoints can be inferred from the geometry of the scene, and can be therefore generated via action labels or language. An example of automatic KP generation can be seen in Fig. 4. Given a point cloud of the scene with semantic labels and a language description of a task, we can use simple heuristics to generate keypoints that can synthesize the described motion. More details can be found in the supp. mat.

#### 3.3. From Action Keypoints to an Anchor Pose

The first step is to infer a pose from the action keypoints in a target location  $\mathbf{a} = \{\mathbf{k}_i\}_{i=1}^P$ , where  $\mathbf{k}_i \in \mathbb{R}^3$  indicates the desired locations for corresponding SMPL joints denoted as  $m_i(\cdot)$ . We find as few as three to four joints ( $P = 3, 4$ ) are usually sufficient. Since the problem is heavily under-constrained we optimize the latent space of VPOSER [50] denoted as  $\mathbf{z}$ . Denoting  $f(\mathbf{z}) \mapsto (\phi, \theta)$  as the mapping from the latent space  $\mathbf{z}$  to the SMPL pose parameters, we minimize the following objective

$$\mathbf{z}, \mathbf{t} = \arg \min_{\mathbf{z}, \mathbf{t}} \sum_{i=1}^P \|m_i(f(\mathbf{z}), \mathbf{t}) - \mathbf{k}_i\|_2 \quad (1)$$

Please see the supplementary material for further details to make the optimization well behaved. We repeat this step for each target action  $\mathbf{a}_1 \dots \mathbf{a}_N$ , obtaining  $N$  pose-anchors  $\mathcal{A} = \{\mathbf{t}_i^A, \phi_i^A, \theta_i^A\}_{i=1:N}$ .

### 3.4. Canonical Coordinate Frame

One of our key ideas to synthesize motion in 3D scenes is to make transformers synthesize motion that always converge at the origin of a canonical coordinate frame. This way at test time long motion is composed by consecutively going to the next goal placed at the origin. Thus, we canonicalize the training sequence clips by using the planar translation  $\mathbf{t}_C$ , and rotation  $\mathbf{R}_C$  of the last frame in a sequence clip as follows

$$\phi_j^C = \mathbf{R}_C^{-1} \phi_j, \mathbf{r}_j^C = \mathbf{R}_C^{-1}(\mathbf{r}_j - \mathbf{t}_C). \quad (2)$$

By construction, this transformation outputs a new set of  $L$  frames  $[(\mathbf{r}^C, \phi^C, \theta)_j]_{j=1:L}$ , where the last pose is at the origin and oriented towards a canonical axis of orientation  $\gamma$  (arbitrary fixed axis). Let  $\mathbf{X}$  denote a matrix whose columns are vectorized motion parameters (pose and translation combined) We will use the following notation to denote the canonicalization in Eq. (2) for a full sequence as

$$\mathbf{X}^C = C(\mathbf{X}; \mathbf{R}_C, \mathbf{t}_C) \quad (3)$$

Synthesizing motion in the goal-centric canonical coordinate frame, allows us to synthesize walking motion along paths in a 3D scene (Sec. 3.5) and transitions in and out of actions (Sec. 3.6) without the need for scene registered data.

### 3.5. WalkNet

**Training.** Using walking sequence clips of variable length  $L$  canonicalized (last pose at origin), we train *WalkNet*. *WalkNet* takes  $K$  motion frames as input  $\mathcal{W}_{inp} = [(\mathbf{r}^W, \phi^W, \theta^W)_j]_{j=1:K}$  and predicts the next  $K$  frames in the sequence  $\mathcal{W}_{out} = [(\mathbf{r}^W, \phi^W, \theta^W)_j]_{j=K+1:2K}$ . The training sub-clips of size  $2K < L$  are randomly sampled from the training walking sequences.

Expressing sequences as matrices (columns are translations and poses) as explained in the previous section, the transformer takes as input a matrix  $\mathbf{X}_{in} \in \mathbb{R}^{K \times 219}$  and outputs a matrix  $\mathbf{X}_{out} \in \mathbb{R}^{K \times 219}$ . We denote the learned mapping as  $T : \mathbf{X}_{in} \mapsto \mathbf{X}_{out}$ . Note that we input the pose as vectorized joint rotation matrices, which make learning more stable compared to using joint angles.

**Test time.** We use the *WalkNet* to follow long paths, by breaking the path into intermediate goals that are canonicalized to the origin (Fig. 3). To traverse scenes avoiding obstacles, we compute the path between the seed motion  $\mathcal{I}$  and the first anchor pose  $\mathcal{A}_1$  using A-star[24]. Along the path, we sample  $P$  goals and compute tangents to the path:

$\{\mathbf{q}_p, \mathbf{l}_p \in \mathbb{R}^3\}_{p=1 \dots P}$ . Then we recursively canonicalize such that tangents  $\mathbf{l}_p$  align with the canonical axis  $\gamma$ . Hence, canonical translation and rotation are computed as follows

$$\mathbf{t}_C = \mathbf{q}_p, \quad \mathbf{R}_C = \exp(\widehat{\mathbf{l}_p \times \gamma}) \quad (4)$$

where  $\exp(\cdot)$  is the exponential map recovering the rotation from the screw-symmetric matrix  $\widehat{\mathbf{l}_p \times \gamma}$ . With this, the motion sequence from goal  $p - 1$  to goal  $p$  is obtained by canonicalizing, predicting future motion with the learned mapping  $T$  and uncanonicalizing

$$\mathbf{X}_{in} \xrightarrow{C(\cdot, \mathbf{R}_C, \mathbf{t}_C)} \mathbf{X}_{in}^C \xrightarrow{T} \mathbf{X}_{out}^C \xrightarrow{C(\cdot, \mathbf{R}_C^T, -\mathbf{t}_C)} \mathbf{X}_{out}. \quad (5)$$

Although the transformer outputs  $K$  future frames, at test time, we use it recursively with a stride of 1 for better performance. That means we effectively predict one pose at a time, and we discard the  $K + 1 : 2K$  frames. In this manner, the motion always goes to the origin, we never have to explicitly send the goal coordinates as input to the network, and we do not drift. When we are sufficiently close to an anchor pose, we predict the transition with *TransNet*.

### 3.6. TransNet

We synthesize transitions between walks and actions again in a canonicalized frame. To do so, we train *TransNet* - a transformer based motion inbetweenner - using AMASS sequences placed in the canonical coordinate frame. The task of *TransNet* is to fill in the motion from a seed sequence  $\mathbf{X}_{in}$  to a target *anchor pose*.

**Training.** We train *TransNet* by asking it to recover training clips from masked out ones. We observe that directly asking to infill many frames does not work reliably. Inspired by training of language models [13], we progressively grow the mask during training until the desired length. Formally, let  $\mathbf{X}$  be a training clip of length  $M$ , let  $\mathbf{V} \in [0, 1]^{M \times 219}$  be a matrix mask with zero-column vectors for the frames that need to be infilled. The network is tasked to recover  $\mathbf{X}$  from the masked out matrix  $\mathbf{X} \odot \mathbf{V}$ . The mask  $\mathbf{V}$  is progressively grown to mask all the motion frames between  $\frac{M}{2}$  to  $M - 1$  frames - everything except the seed motion and the last anchor pose. For more details, please see supp. mat.

**Test time** We use *TransNet* to synthesize transitions in 3D scenes by moving  $\frac{M}{2}$  frames of a motion sequence into the canonical coordinate frame by using the orientation and position of the motion-anchor pose - the motion-anchor pose is thus placed at the origin of the canonical coordinate frame. *TransNet* is then tasked to infill the missing frames. (Fig. 5)

### 3.7. Chained actions

With our models and representations we can chain actions trivially. At run time, we have to satisfy an arbitrary

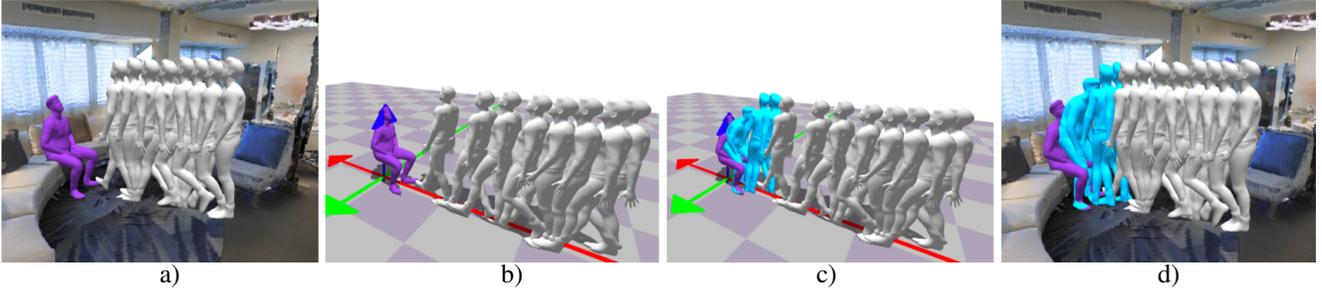


Figure 5. Using a) the motion-anchor pose in the 3D scene (purple), b) we move the motion sequence into the canonical coordinate frame. c) There *TransNet* synthesizes transitions (blue) between the input motion and the pose placed at the origin (purple). d) Once the motion is synthesized, we move it back to the scene coordinate frame.

number of actions keypoints  $\{\mathbf{a}_1, \dots, \mathbf{a}_N\}$  at different locations. First we compute anchor poses as explained in Sec. 3.3. Obstacle free paths connecting the locations of actions are computed with  $A^*$ . We rely on *WalkNet* to follow paths until we are sufficiently close to the first anchor pose. Feeding *TransNet* with the last  $M/2$  predicted frames of *WalkNet* and the anchor pose, we predict the transition into the first anchor pose. To transition out we also use *TransNet* with no modification. We sample a location along the path from  $\mathbf{a}_1$  to  $\mathbf{a}_2$  at a fixed distance  $\delta$  and place a walking pose from our database. *TransNet* then can transition into this walking pose (Fig. 5). Then we activate *WalkNet* and the process is repeated until all actions are executed. In addition, we can repeatedly use *TransNet* to execute several actions at the same location, like grabbing at different heights.

## 4. Experiments

In this section we present implementation details of our method. Next, we compare our approach with existing methods. Our experiments show that we clearly outperform existing baselines. Next, we ablate our design choices and finally present qualitative results of our method.

### 4.1. Implementation Details

*WalkNet* and *TransNet* are BERT [13] style full-attention transformers. Both consist of 3 attention layers - each composed of 8 attention heads. We use an embedding size of 512 for both transformers. For more details please see the supplementary material. For training both transformers, we set the learning rate to  $1e^{-5}$ . Both networks are trained using an  $L2$  loss. We set  $M = 120$  and  $K = 30$ . We experimented with three different values of  $M$  and found that  $M = 120$  produces the least foot-skating. Please see the supplementary material for this experiments.

### 4.2. Datasets

**Motion Data:** To train *TransNet* and *Walknet* we use the large mocap dataset AMASS [46]. For exact details how this is done, please see the supplementary material.

**Scene Datasets:** We demonstrate that our method is able to generate realistic human motion in scenes from **Matterport3D**, **HPS**, **Replica** and **ScanNet** datasets. All these datasets have been reconstructed using RGB-D scanners or LIDAR scanners and contain scans with sizes ranging from  $20 m^2$  to  $1000 m^2$ . While Replica, Matterport scenes contain perfect geometry, ScanNet scenes do not. Our method is able to generalize across all these scenes.

### 4.3. Evaluation Metrics:

We compare our method with existing baselines using perceptual studies and a foot skate metric. Additionally, we ablate various components of our method with the same foot skate metric.

**Perceptual Study:** We synthesize two motion sequences - one using our method and another using a baseline method and show the two synthesized sequences to participants in our perceptual study. The participant is asked to answer “Which motion looks most realistic?” and “Which motion satisfies scene constraints best?”. The study is conducted in such a manner that the participant is forced to choose one of two motions in front of him.

**Foot Skating (FS):** The foot-skate metric measures how much foot-skate occurs during a synthesized motion measured in cm/frame. For  $N$  frames, it is defined as:

$$s = \sum_{p=1}^N [v_p (2 - 2^{\frac{h_p}{H}}) \mathbf{1}_{h_p <= H}]$$

where  $h_p$  is the height of the vertex and  $v_p$  is the velocity of a foot vertex on the right toe in frame  $p$  and  $H = 2.5$  cm

### 4.4. Comparison with Baselines

As aforementioned, no method addresses the task of continual motion synthesis in arbitrary 3D scenes. For completeness we do our best to compare our approach with three existing methods: SAMP [26], GAMMA [75], Wang et al. [66] which all generate animator guided motion by navigating  $A^*$  paths in 3D scenes. Though, these methods use different forms of animator guidance - such as action labels,

	Ours	SAMP	Ours	GAMMA	Ours	Wang et al.
Which motion is most realistic (%) $\uparrow$	<b>71.8</b>	28.2	<b>95.6</b>	4.4	<b>100</b>	0
Which motion satisfies scene constraints best (%) $\uparrow$	<b>76.8</b>	23.2	<b>100</b>	0	<b>100</b>	0

Table 1. Comparisons between our method and existing baselines using a perceptual study.

	Language	Manual	<i>WalkNet</i>	MoGlow	<i>TransNet</i>	NeMF
Foot Skate (cm/f) $\downarrow$	0.93	<b>0.92</b>	<b>0.91</b>	1.88	<b>1.1</b>	1.54
User Study (%) $\uparrow$	<b>53.8</b>	46.2	<b>75.7</b>	24.3	<b>66.8</b>	33.2

Table 2. Analysis of different components in our method. We compare our method with different baselines across three design components: using language based or manually specified keypoints, the walking motion and the transition motion.

	Ours	SAMP	GAMMA	Wang et al.
Foot-skate $\downarrow$	<b>0.91</b>	1.34	0.94	4.53

Table 3. Comparisons between our method and existing baselines using the foot-skate metric.

we modify them by incorporating the KP information used by our method. Note that except GAMMA, none of these baselines can be deployed in arbitrary 3D scenes without significant modifications, as described below.

**SAMP:** SAMP is written entirely in Unity and can only synthesize sitting and lying actions in synthetic scenes. Unlike SAMP our method requires no manual action annotation. The object of interaction and the action to perform are the animator guidance provided as input to SAMP. SAMP synthesizes motion by explicitly conditioning on the geometry of the object of interaction, and by navigating A\* paths. For comparison with SAMP, we represent the object of interaction in one of our test scenes with a synthetic object in Unity. Using KPs, we represent the orientation of action in our test scene and use this orientation to port A\* paths used in our test 3D scene into Unity and run the publicly available code of SAMP. For exact details, please see the supp. mat. Note that SAMP cannot synthesize chained actions nor can it be deployed in arbitrary 3D scenes. For instance it cannot sit on stairs nor can it perform a grabbing action near a bookshelf. The comparison is included for completeness as SAMP also navigates A\* paths.

**Wang et al.:** We run the pre-trained code of Wang et al. on scenes from HPS, Replica and Matterport Datasets. Instead of using action labels to generate anchor poses as done in the original paper, we replace this step with the motion anchors generated using our inverse kinematics step. Since Wang et al. [66] is trained using the PROX dataset and synthesizes navigational motion across A\* paths by explicitly conditioning on scene geometry, it does not generalize at all to 3D scenes beyond these datasets.

**GAMMA:** GAMMA only navigates 3D scenes and is unable to synthesize human-scene interaction. Similar to the navigation part of our method, it uses the start and end of a path as animator guidance. For the purpose of this comparison, we generate a set of paths in 3D scenes using A\* and synthesize walking motion along this path using

GAMMA and our method. GAMMA is unable to follow the exact waypoints of the path and as such produces significant interpenetrations with the 3D scene.

For visualizations of motion synthesized by these baselines, please see the supplementary video. We synthesize 5 motion sequences of a total duration of 300 seconds using each method in 5 different scenes for our perceptual study. In Tab. 1, we report the results of our perceptual study with 50 participants (see Sec. 4.3). Each column corresponds to the percentage of users who choose the method corresponding to the column heading. Our results are preferred by a vast majority of the participants. In Tab. 3, we report the numbers corresponding to foot skate metric.

#### 4.5. Ablation Studies

##### Can *TransNet* be replaced with other inbetweeners?

We compare *TransNet* with the SoTA inbetweening method NeMF [29] for the task of transitioning in and out of actions. For our task of infilling  $\frac{M}{2} - 1$  frames in the canonical coordinate frame, *TransNet* produces more natural motion and less foot skating. We hypothesize that this occurs as NeMF is a general purpose inbetweenner that can infill an arbitrary number of frames, whereas *TransNet* is a motion inbetweenner custom designed for the purpose of infilling  $\frac{M}{2} - 1$  motion frames in the canonical coordinate frame. We conduct a new user study with 36 participants, asking users to rate the naturalness of 20 motion sequences by NeMF and *TransNet*. Results are reported in Tab 2.

##### Can *WalkNet* be replaced with other path following methods?

We provide comparisons with SAMP, Wang et al, and GAMMA which all navigate A\* paths. As our experiments illustrate, our method outperforms these existing methods for navigation. For further completeness, we trained the SoTA walking method, MoGlow [30], on our walking data. When deployed on 150-200 meter long A\* paths, it produces significant foot-skating after about 30 secs. We hypothesize that this occurs because MoGlow synthesizes motion in an egocentric coordinate frame, and hence the control signal provided by A\* changes rapidly which leads MoGlow to synthesize motion with significant drift. We compare our method to MoGlow on these paths

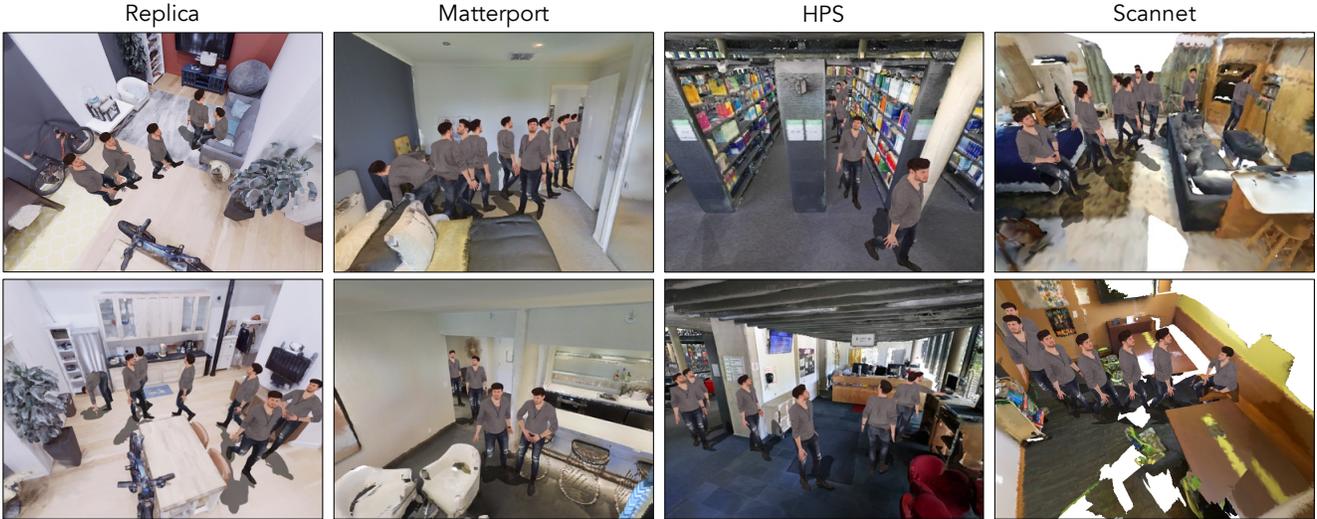


Figure 6. Our method allows to generate motion that generalizes across different scenes. Here we show motion generation in scenes from 4 different datasets: Replica [61], Matterport [8], HPS [22] and Scannet [10].



Figure 7. The keypoint representation allows us to generate diverse and highly controllable motion. We show here examples of different grabbing, sitting and newly defined motions.

using a user study with 36 participants in Tab. 2, where our approach outperforms MoGlow.

**How well does language based keypoint placement work?** In this experiment, we compare motion synthesized using manual keypoint placement with language based keypoint placement. We synthesize 5 motion sequences using keypoints generated by these two approaches and compare the synthesized sequences using a user study with 36 participants. When used for motion synthesis, these KPs produce similar quality as manual KP placement (Tab. 2).

**How long does it take for a user to provide keypoints manually?** We develop a user interface which allows users to navigate 3D scenes and to click on locations of interaction. We instruct 7 participants how to navigate 3D scenes with our user interface. On average it takes 245 seconds for users to learn the interface. We then ask each user to provide 5 sets of 3 action keypoints (the location of the root and the two feet or the location of one hand and two feet) for a total of 15 keypoints per scene in 5 different 3D scenes. On average it takes 125 seconds to select these points per scene.

#### 4.6. Qualitative Results

Please watch the supp. video for qualitative evaluation. In Figure 6, we demonstrate examples of motion generated in scenes from 4 different datasets: Replica [61], Matterport[8], HPS[22] and Scannet[10]. Moreover, representing the motion as Action Keypoints allows us to have high control and diversity over the generated motions. In Figure 7 we show how this representation allows us to sit or pick objects at different heights (left column), or generate actions such as grabbing with two hands or stretching.

#### 5. Limitations and Conclusions

We presented the first method to synthesize continual human motion in scenes from the HPS, Matterport, ScanNet, and Replica. Our core contribution is a novel method for long-range motion synthesis via iterative canonicalization and the use of keypoints to decouple scene reasoning from motion synthesis, and provide a flexible interface to synthesize motion. We demonstrated that our method works better than existing solutions that generate motion in 3D scenes.

While our approach presents an important step towards long-range motion synthesis in 3D scenes, it also has limitations: It assumes a horizontal floor and thus does not support scenes with uneven floors. It also assumes valid keypoint configurations: if the keypoints provided by the user do not conform to a valid pose, the pose produced by the IK step will not look realistic, producing unnatural motion. In the future we hope to remove this limitation by reducing the number of required keypoint inputs. We hope that the proposed approach drives new research towards continual human motion in arbitrary 3D scenes.

## References

- [1] Aksan, E., Cao, P., Kaufmann, M., Hilliges, O.: A spatio-temporal transformer for 3d human motion prediction (2020) [2](#), [3](#)
- [2] Aksan, E., Kaufmann, M., Hilliges, O.: Structured prediction helps 3d human motion modelling. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2019), first two authors contributed equally. [3](#)
- [3] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 961–971 (2016) [2](#)
- [4] Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15935–15946 (2022) [3](#)
- [5] Brand, M., Hertzmann, A.: Style machines. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. p. 183–192. SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., USA (2000) [2](#)
- [6] Cao, Z., Gao, H., Mangalam, K., Cai, Q.Z., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: ECCV (2020) [3](#)
- [7] Cao, Z., Gao, H., Mangalam, K., Cai, Q., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: ECCV (2020) [3](#)
- [8] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. International Conference on 3D Vision (3DV) (2017) [2](#), [3](#), [8](#)
- [9] Clavet, S.: Motion matching and the road to next-gen animation. In: Game Development Conference (2016) [3](#)
- [10] Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2017) [2](#), [8](#)
- [11] Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021) [2](#)
- [12] Delaitre, V., Fouhey, D.F., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 284–298. Springer Berlin Heidelberg, Berlin, Heidelberg (2012) [3](#)
- [13] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) NAACL-HLT (1). Association for Computational Linguistics (2019) [5](#), [6](#)
- [14] Duan, Y., Shi, T., Zou, Z., Lin, Y., Qian, Z., Zhang, B., Yuan, Y.: Single-shot motion completion with transformer (2021) [3](#)
- [15] Fang, A.C., Pollard, N.S.: Efficient synthesis of physically valid human motion. ACM Transactions on Graphics (TOG) **22**(3), 417–426 (2003) [3](#)
- [16] Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People watching: Human actions as a cue for single view geometry. International journal of computer vision **110**(3), 259–274 (2014) [3](#)
- [17] Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4346–4354 (2015) [2](#)
- [18] Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: CVPR 2011. pp. 1529–1536. IEEE (2011) [3](#)
- [19] Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013) [2](#)
- [20] Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object

- perception. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007) **3**
- [21] Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR 2011. pp. 1961–1968. IEEE (2011) **3**
- [22] Guzov\*, V., Mir\*, A., Sattler, T., Pons-Moll, G.: Human positioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2021) **2, 3, 8**
- [23] 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) (September 2017) **2**
- [24] Hart, P., Nilsson, N., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. IEEE Transactions on Systems Science and Cybernetics **4**(2), 100–107 (1968). <https://doi.org/10.1109/tssc.1968.300136>, <https://doi.org/10.1109/tssc.1968.300136> **5**
- [25] Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.: Robust motion in-betweening **39**(4) (2020) **3**
- [26] 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 (Oct 2021) **2, 3, 6**
- [27] Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: Proceedings International Conference on Computer Vision. pp. 2282–2292. IEEE (Oct 2019), <https://prox.is.tue.mpg.de> **2, 3**
- [28] Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3D scenes by learning human-scene interaction. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 14708–14718 (Jun 2021) **3**
- [29] He, C., Saito, J., Zachary, J., Rushmeier, H., Zhou, Y.: Nemf: Neural motion fields for kinematic animation. In: NeurIPS (2022) **7**
- [30] Henter, G.E., Alexanderson, S., Beskow, J.: Moglow: Probabilistic and controllable motion synthesis using normalising flows. ACM Transactions on Graphics (TOG) **39**(6), 1–14 (2020) **7**
- [31] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997) **2**
- [32] Holden, D., Kanoun, O., Perepichka, M., Popa, T.: Learned motion matching. ACM Transactions on Graphics (TOG) **39**(4), 53–1 (2020) **3**
- [33] Holden, D., Komura, T., Saito, J.: Phase-functioned neural networks for character control. ACM Transactions on Graphics (TOG) **36**(4), 1–13 (2017) **3**
- [34] Kaufmann, M., Aksan, E., Song, J., Pece, F., Ziegler, R., Hilliges, O.: Convolutional autoencoders for human motion infilling. In: 2020 International Conference on 3D Vision (3DV). pp. 918–927 (2020). <https://doi.org/10.1109/3DV50981.2020.00102> **3**
- [35] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. ICLR (2014) **2**
- [36] Kipf, T.N., Welling, M.: W. 2016. semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations **2**
- [37] Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. In: ACM SIGGRAPH 2008 classes, pp. 1–10 (2008) **3**
- [38] Lee, J., Chai, J., Reitsma, P.S., Hodgins, J.K., Pollard, N.S.: Interactive control of avatars animated with human motion data. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques. pp. 491–500 (2002) **3**
- [39] Li, J., Yin, Y., Chu, H., Zhou, Y., Wang, T., Fidler, S., Li, H.: Learning to generate diverse dance motions with transformer. arXiv preprint arXiv:2008.08171 (2020) **3**
- [40] 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) **2**
- [41] Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Learn to dance with aist++: Music conditioned 3d dance generation. In: ICCV (2021) **3**
- [42] 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 Conference on Computer Vision and Pattern Recognition. pp. 12368–12376 (2019) **3**

- [43] Ling, H.Y., Zinno, F., Cheng, G., van de Panne, M.: Character controllers using motion vaes. *ACM Trans. Graph.* **39**(4) (2020) [3](#)
- [44] Liu, C.K., Hertzmann, A., Popović, Z.: Learning physics-based motion style with nonlinear inverse optimization. *ACM Transactions on Graphics (TOG)* **24**(3), 1071–1081 (2005) [2](#)
- [45] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics* (2015) [4](#)
- [46] Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: *International Conference on Computer Vision*. pp. 5442–5451 (Oct 2019) [2, 6](#)
- [47] Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 9489–9497 (2019) [2](#)
- [48] Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2891–2900 (2017) [2](#)
- [49] Oreshkin, B.N., Valkanas, A., Harvey, F.G., Ménard, L.S., Bocquet, F., Coates, M.J.: Motion inbetweening via deep  $\delta$ -interpolator (2022) [3](#)
- [50] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2019) [4](#)
- [51] Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: *International Conference on Computer Vision (ICCV)*. pp. 10985–10995 (October 2021) [3](#)
- [52] Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: *European Conference on Computer Vision (ECCV)* (2022) [3](#)
- [53] Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., Torralba, A.: Virtualhome: Simulating household activities via programs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8494–8502 (2018) [3](#)
- [54] Reitsma, P.S., Pollard, N.S.: Evaluating motion graphs for character animation. *ACM Transactions on Graphics (TOG)* **26**(4), 18–es (2007) [3](#)
- [55] Safonova, A., Hodgins, J.K.: Construction and optimal search of interpolated motion graphs. *ACM Transactions on Graphics (SIGGRAPH 2007)* **26**(3) (Aug 2007) [3](#)
- [56] Safonova, A., Hodgins, J.K., Pollard, N.S.: Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Transactions on Graphics (ToG)* **23**(3), 514–521 (2004) [3](#)
- [57] Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: PiGraphs: Learning Interaction Snapshots from Observations. *ACM Transactions on Graphics (TOG)* **35**(4) (2016) [3](#)
- [58] Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. *ACM Trans. Graph.* **38**(6) (Nov 2019) [2, 3](#)
- [59] Starke, S., Zhao, Y., Komura, T., Zaman, K.: Local motion phases for learning multi-contact character movements. *ACM Trans. Graph.* **39**(4) (Jul 2020) [3](#)
- [60] Starke, S., Zhao, Y., Zinno, F., Komura, T.: Neural animation layering for synthesizing martial arts movements. *ACM Trans. Graph.* **40**(4) (Jul 2021) [3](#)
- [61] Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briaies, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., Nardi, R.D., Goesele, M., Lovegrove, S., Newcombe, R.: The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797* (2019) [2, 8](#)
- [62] Taheri, O., Choutas, V., Black, M.J., Tzionas, D.: Goal: Generating 4d whole-body motion for hand-object grasping. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13263–13273 (2022) [2, 3](#)
- [63] 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), <https://grab.is.tue.mpg.de> [3](#)
- [64] Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence* **30**(2), 283–298 (2007) [2](#)

- [65] Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9401–9411 (2021) [2](#), [3](#)
- [66] Wang, J., Rong, Y., Liu, J., Yan, S., Lin, D., Dai, B.: Towards diverse and natural scene-aware 3d human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20460–20469 (2022) [2](#), [3](#), [6](#), [7](#)
- [67] Wang, J., Yan, S., Dai, B., Lin, D.: Scene-aware generative network for human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12206–12215 (2021) [3](#)
- [68] Wang, X., Girdhar, R., Gupta, A.: Binge watching: Scaling affordance learning from sitcoms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2596–2605 (2017) [3](#)
- [69] Wu, Y., Wang, J., Zhang, Y., Zhang, S., Hilliges, O., Yu, F., Tang, S.: Saga: Stochastic whole-body grasping with contact. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022) [2](#), [3](#)
- [70] Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., Kautz, J.: Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) [3](#)
- [71] Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: European Conference on Computer Vision. pp. 346–364. Springer (2020) [2](#)
- [72] 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) [3](#)
- [73] Zhang, X., Bhatnagar, B.L., Starke, S., Guzov, V., Pons-Moll, G.: Couch: Towards controllable human-chair interactions (October 2022) [2](#), [3](#)
- [74] Zhang, Y., Black, M.J., Tang, S.: We are more than our joints: Predicting how 3D bodies move. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). p. 3372–3382 (Jun 2021) [2](#), [3](#)
- [75] Zhang, Y., Tang, S.: The wanderings of odysseus in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20481–20491 (2022) [3](#), [6](#)
- [76] Zhao, K., Wang, S., Zhang, Y., Beeler, T., Tang, S.: Compositional human-scene interaction synthesis with semantic control. In: European conference on computer vision (ECCV) (2022) [3](#)