NIFTY: Neural Object Interaction Fields for Guided Human Motion Synthesis

Nilesh Kulkarni[1] Davis Rempe[2] Kyle Genova[3] Abhijit Kundu[3]
Justin Johnson[1] David Fouhey[4] Leonidas Guibas[3,5]
University of Michigan[1] NVIDIA[2] Google[3] New York University[4] Stanford University[5]

Abstract

*We address the problem of generating realistic 3D motions of humans interacting with objects in a scene. Our key idea is to create a neural interaction field attached to a specific object, which outputs the distance to the valid interaction manifold given a human pose as input. This interaction field guides the sampling of an object-conditioned human motion diffusion model, so as to encourage plausible contacts and affordance semantics. To support interactions with scarcely available data, we propose an automated synthetic data pipeline. For this, we seed a pre-trained motion model, which has priors for the basics of human movement, with interaction-specific anchor poses extracted from limited motion capture data. Using our guided diffusion model trained on generated synthetic data, we synthesize realistic motions for sitting and lifting with several objects, outperforming alternative approaches in terms of motion quality and successful action completion. We call our framework **NIFTY**: **N**eural **I**nteraction **F**ields for **T**rajectory s**Y**nthesis. NIFTY results are available on* https://nileshkulkarni.github.io/nifty

1. Introduction

Predicting human-object interaction motions is a key component of systems for vision, robotics, and animation. For example, consider generating the motion of a person sitting on a chair or table as in Fig. 1. Achieving this in general requires progress on several important independent sub-problems: (i) navigating in the scene while avoiding obstacles, (ii) approaching the object and initiating an interaction through contact, and (iii) potentially moving the object and transitioning to subsequent actions. In this work, we address the second problem of approaching and initiating contact, which is the core of the interaction. We term this problem the "*last mile*" of interaction. Unlike navigation in stage (i), which is primarily collision avoidance, solving the last mile requires accounting for object affordances and changes in contacts that influence human motion. Afterwards, kinematic [8] or learning-based methods [56] can be used in stage (iii) to update object pose during continued motion.

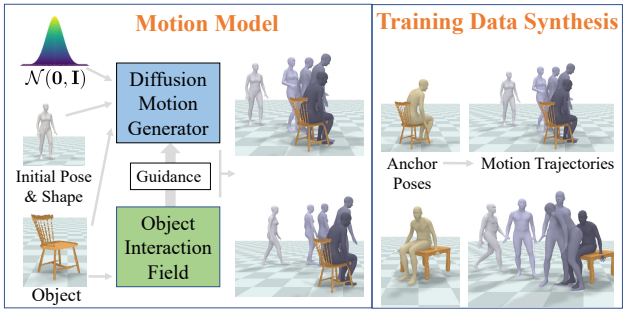

Figure 1. **NIFTY Overview.** (Left) Our learned **object interaction field** guides an object-conditioned **diffusion model** during sampling to generate plausible human-object interactions like sitting. (Right) Our automated *training data synthesis pipeline* generates data for this model by combining a scene-unaware motion model with small quantities of annotated interaction anchor pose data.

One challenge of generating motion in the last mile is developing an *effective model*. Recent generative human motion models [15, 34, 45] synthesize realistic movements but are unaware of scene context. To address this, some approaches condition motion synthesis on entire scene geometry, *e.g.*, a scanned point cloud [20, 50, 52, 53]. However, these methods can be multi-stage and rely on post-processing optimization [50, 52] or hand-designed guidance constraints [20] to avoid physical artifacts like scene penetration and foot sliding [53]. Other approaches are conditioned on a single object [13, 41, 61] and focus on a small set of actions (*e.g.*, sitting specifically on a chair), but make action-specific modeling assumptions that require per-frame annotations for contact, action class, and phase [41].

A second major challenge of learning motion for the last mile is the lack of reliable *paired human-object data*. Datasets that capture humans in general scenes offer a variety of motions but limited quality due to the complexity of capture [11]. Meanwhile, specific interactions with single objects can be captured with higher quality [3, 12, 42, 61] but are of limited size and scope. Methods that train on such object data require extensive data augmentation [12, 41, 61], and extending this detailed capture setup to new objects and motions is difficult and expensive.

In this work, we tackle both the *modeling* and *data* challenges to enable generating realistic interactions with a variety of objects, such as sitting on a chair, table, or stool and lifting a suitcase, chair, *etc*. In terms of modeling, we extend a human motion diffusion model [45] to condition on object geometry and propose guiding generation with a data-driven *human-object interaction field*. This interaction field is object-centric, taking a human pose as input and learning to regress the distance to a valid interaction pose (*e.g*., the final sitting pose). The field forms a guidance loss whose gradient pushes the test-time denoising process to produce high-quality interaction motions while accounting for both physical and semantic factors, unlike prior guidance approaches [20]. We refer to this framework as **NIFTY**: **N**eural **I**nteraction **F**ields for **T**rajectory s**Y**nthesis.

To train this model and overcome the lack of available mocap data, we develop an *automated data pipeline* that leverages a powerful pre-trained and scene-agnostic human motion model. Starting from an anchor pose that captures the moment of contact (*e.g*., the final sitting pose in Fig. 1, right), the pre-trained motion model is used to sample a large variety of motions that *end* in the anchor pose. As shown in Fig. 1, our interaction field, diffusion model, and motion data pipeline make up a general framework to synthesize human-object interactions for a desired character that is flexible to multiple actions, even when dense mocap data is unavailable.

We evaluate NIFTY on sitting and lifting interactions for a variety of objects. In our user study, NIFTY's motions are preferred over 80% of the time to state-of-the-art baselines [12, 53], while outperforming on quantitative interaction metrics. Overall, this work contributes **(1)** a novel object interaction field to guide an object-conditioned motion diffusion model to synthesize realistic interactions, **(2)** an automated synthetic data generation technique to produce many interactions from limited pose data, and **(3)** high-quality motion synthesis for interactions with several objects.

2. Related Work

Synthesizing Human Motion and Interactions. Unlike works that generate human motion in isolation [15, 18, 32, 34, 45, 62], our work focuses on incorporating environmental context [4, 5]. Some approaches condition motion generation on scanned scenes that encompass multiple objects [20, 50, 51, 52], but the limited availability of paired scene-motion data hurts the quality of results. Object-centric models like NSM [41], SAMP [13], and COUCH [61] generate higher-quality motions by narrowing the problem scope to specific single-object interactions, *e.g*., generating motions for a single character [13, 14, 41] or a limited set of actions [61] like sitting on a chair. These models rely on high-quality motion capture datasets, and some require additional dense annotations like contacts [41, 61], actions [12], and phases [41]. DIMOS [63] avoids needing paired motion-object data by learning a control policy in a motion prior latent space, but the policy requires a full-body goal pose as input at test time. Concurrent work ROAM [60] uses a neural descriptor field (NDF) [38] to optimize an interaction goal pose that becomes the input for a separate autoregressive motion model. Our work focuses on interactions with individual objects, but utilizes diffusion guidance with a learned interaction field to minimize artifacts and avoid the need for a goal pose input at runtime. We use a novel data generation pipeline to learn such interactions from limited data.

Motion Diffusion Models. Following success in other domains [16, 17, 21, 30, 35, 40, 65], diffusion models have enabled generating high-quality, full-body 3D human motion [6, 45, 47, 59]. While some models enable controllability through kinematic pose constraints [24, 37, 55], they are unaware of scene context. SceneDiffuser [20] generates human motion conditioned on a scene point cloud after training on noisy paired scene-motion data [11]. It employs gradient-based guidance with hand-designed objectives to encourage collision-free, contact-driven, and smooth motion during denoising. InterDiff [56] learns an interaction corrector that intermittently updates human and object motion predictions during the denoising process if contact and penetration constraints are violated. Our interaction field is *data-driven* and guides denoising *at every step* to generate plausible motion that approaches an object and ends in valid contact.

Neural Distance Fields for Pose. Grasping Fields [23] parameterize hand-object grasping through a spatial field that outputs distances to valid hand-object grasps. PoseNDF [46] learns an object-unaware distance field in the full-body pose space for human poses. NGDF [54] and SE(3)-DiffusionFields [48] learn a field in the robot gripper pose space to define a manifold of valid object grasps. Our object interaction field extends this idea to full-body human-object interactions by learning to predict the distance between a human pose and the interaction pose manifold. Unlike prior works, we use this field to guide denoising.

Human Interaction Data. Learning human-object interactions is hampered by the challenge of capturing humans in scenes. Datasets that contain full scenes paired with human motion [10, 11, 19, 36, 64] are relatively small and often noisy due to capture difficulties. Single-object interaction datasets with a small set of objects [3, 13, 22, 42, 61] are better quality but are small with limited scope. Recent approaches use automated synthetic data generation, *e.g*., 3D scenes can be inferred from pre-recorded human motions to get plausible paired scene-motion data [53, 57, 58]. However, motions from these methods are limited to available pre-recorded mocap. Our data generation requires a small set of interaction anchor poses and generates novel motions not contained in prior datasets using tree-based rollouts [62] from a pre-trained generative model [34].

3. Method

In this section, we detail our NIFTY pipeline for learning to synthesize realistic human-object interaction motions. §3.1 introduces a conditional diffusion model to generate human motions given the geometry of an object. §3.2 details the object-centric interaction field to guide the diffusion model at sampling time and §3.3 discusses the synthetic data generation process to train the model.

3.1. Motion Generation through Diffusion

Motion Representation. Motion generation is formulated as predicting a sequence of 3D human pose states. Our pose representation is based on the SMPL body model [25] and is similar to prior successful human motion diffusion models [9, 45]. Pose X_i at frame i in a motion sequence is:

$$X_i = \{\boldsymbol{j}_i^p, \boldsymbol{j}_i^r, \boldsymbol{j}_i^v, \boldsymbol{j}_i^\omega, \boldsymbol{t}_i^p, \boldsymbol{t}_i^v\}, \tag{1}$$

which includes joint positions $\boldsymbol{j}_i^p \in \mathbb{R}^{3\times 22}$, rotations $\boldsymbol{j}_i^r \in \mathbb{R}^{6\times 22}$, velocities $\boldsymbol{j}_i^v \in \mathbb{R}^{3\times 22}$, and angular velocities $\boldsymbol{j}_i^\omega \in \mathbb{R}^{3\times 22}$ for all 22 SMPL joints including the root (pelvis). Additionally, the SMPL global translation $\boldsymbol{t}_i^p \in \mathbb{R}^3$ and velocity $\boldsymbol{t}_i^v \in \mathbb{R}^3$ are included. A motion is a sequence of N poses denoted as $\boldsymbol{\tau} = \{X_1, \ldots, X_N\}$ where all poses are in a canonicalized coordinate frame based on the first pose X_1.

Model Formulation. The diffusion model simultaneously generates all human poses in a motion sequence [45] to achieve a desired interaction. Intuitively, diffusion is a noising process that converts clean data into noise. We want our motion model to learn the reverse of this process so that realistic motions can be generated from randomly sampled noise. Mathematically, forward diffusion is a Markov process with a transition probability distribution:

$$q(\boldsymbol{\tau}^k|\boldsymbol{\tau}^{k-1}) := \mathcal{N}(\boldsymbol{\tau}^k; \mu = \sqrt{1-\beta^k}\boldsymbol{\tau}^{k-1}, \sigma = \beta^k \mathbf{I}), \tag{2}$$

where $\boldsymbol{\tau}^k$ denotes the motion trajectory at the k^{th} noising step, and a fixed β^k is chosen such that $q(\boldsymbol{\tau}^K) \approx \mathcal{N}(\boldsymbol{\tau}^K; \mathbf{0}, \mathbf{I})$ after K steps. Our generative model learns the reverse of this process (denoising), *i.e.* it recovers $\boldsymbol{\tau}^{k-1}$ from a noisy input trajectory $\boldsymbol{\tau}^k$ at each step and doing this repeatedly results in a final clean motion $\boldsymbol{\tau}^0$. Because the model is generating *interaction* motions with an object, we condition denoising on interaction information $C = \{P_o, R_o, \mathbf{b}, X_0\}$, which includes the canonicalized object point cloud $P_o \in \mathbb{R}^{5000\times 3}$, rigid object pose relative to the person $R_o \in \mathbb{R}^{4\times 4}$, SMPL body shape parameters $\mathbf{b} \in \mathbb{R}^{10}$, and starting pose X_0. Each reverse step is then:

$$p_\theta(\boldsymbol{\tau}^{k-1}|\boldsymbol{\tau}^k, C) := \mathcal{N}(\boldsymbol{\tau}^{k-1}; \mu = \mu_\theta(\boldsymbol{\tau}^k, k, C), \sigma = \beta^k \mathbf{I}), \tag{3}$$

where the diffusion step k is also given as input. Instead of predicting the noise $\boldsymbol{\epsilon}^k$ added at each step of diffusion [16, 21], our model M_θ directly predicts the final clean signal $\hat{\boldsymbol{\tau}}^0 = M_\theta(\boldsymbol{\tau}^k, k, C)$ from which the mean $\mu_\theta(\tau^k, k, C)$ is computed [30, 35, 45]. This formulation has the benefit that physically grounded objectives can be easily applied on $\hat{\boldsymbol{\tau}}^0$ in the pose space, which is useful for guidance as discussed below. During training, a ground truth clean trajectory $\boldsymbol{\tau}^0$ is noised and given as input, then the model is trained to minimize the objective $\|\hat{\boldsymbol{\tau}}^0 - \boldsymbol{\tau}^0\|_2^2$.

Sampling and Guidance. At test time, samples are generated from the model given random noise and interaction conditioning C as input. Ensuring that the sampled motions adhere to the geometric and semantic constraints of the object is key to plausible interactions. Diffusion models are well-suited for this, since *guidance* can encourage samples to meet desired objectives at test time [21].

The core of guidance is a differentiable function $G(\boldsymbol{\tau}^0)$ that evaluates how well a trajectory meets a desired objective. This function could be learned [21] or analytic [35]. For our problem, $G(\boldsymbol{\tau}^0)$ evaluates how plausible an interaction motion is for a specific object; §3.2 details how this can be done with a learned object interaction field. Throughout denoising (*i.e.*, sampling), the gradient of the objective function is used to nudge trajectory samples in the correct direction. We use a formulation of guidance that perturbs the clean trajectory output from the model $\hat{\boldsymbol{\tau}}^0$ at every denoising step k as follows [17, 35]:

$$\tilde{\boldsymbol{\tau}}^0 = \hat{\boldsymbol{\tau}}^0 - \alpha \nabla_{\boldsymbol{\tau}^k} G(\hat{\boldsymbol{\tau}}^0) \tag{4}$$

where α controls the guidance strength. The updated trajectory $\tilde{\boldsymbol{\tau}}^0$ is then used to compute μ.

Architecture. As shown in Fig. 2 (right), the denoising model M_θ is based on a transformer encoder-only architecture [45, 49]. Each human pose in the input trajectory $\boldsymbol{\tau}^k$ is a token, while each conditioning in C becomes a separate token. The object point cloud P_o is encoded with a PointNet [33], the rigid pose R_o is encoded with a three-layer MLP, and k is encoded using a positional embedding [44]. Full details are available in the supplementary material.

3.2. Human-Object Interaction Field

After training on human-object interactions, the diffusion model can generate reasonable motion sequences but fails to fully comply with constraints in the last mile of interaction [2, 7], even when conditioned on the object. This causes artifacts such as penetration with the object. To alleviate this issue, we propose to guide motion samples from the diffusion model (Eq. (4)) with a learned objective G that captures realistic interactions for a specific object.

We take inspiration from recent work that uses neural distance fields to learn valid human pose manifolds [46] and robotic grasping manifolds [54]. For our purposes, the field takes in an arbitrary human pose and outputs how far the query pose is from being a "valid" object interaction pose. We define an *interaction pose* to be an *anchor* frame in a

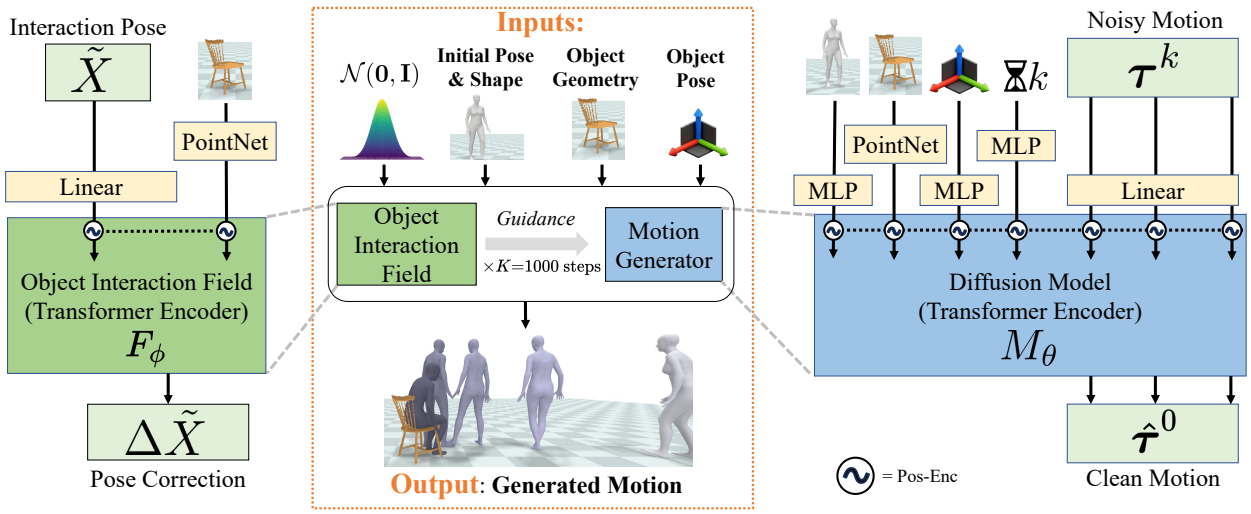

Figure 2. **Model Architecture.** Our full motion synthesis method (middle) consists of an object interaction field F_ϕ (left), which guides the diffusion model M_θ (right) at sampling time to produce plausible interactions. At each step k of denoising, the diffusion model predicts a clean motion $\hat{\boldsymbol{\tau}}^0$ from a noisy motion input $\boldsymbol{\tau}^k$ and conditioning information. The object interaction field takes the last pose from the diffusion output as input, and uses guidance to push the pose towards the valid interaction manifold using a predicted pose correction.

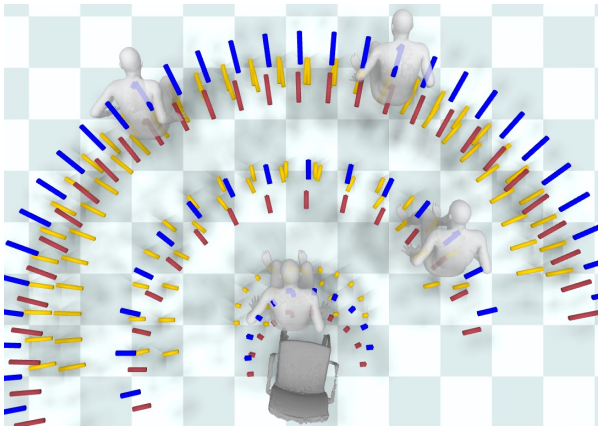

Figure 3. **Interaction Field Visualization.** We query the field in several locations with a sitting pose (a subset shown in grey) and visualize the output for **pelvis**, **feet**, and **neck** joints. All cylinders are oriented towards the chair, indicating the correction vector's magnitude and direction. This correction is due to the misalignment between the sitting pose and chair position.

motion sequence that captures the core of the interaction, *e.g.*, the moment a person settles in a chair after sitting (as in Fig. 1) or contacts an object when lifting.

Our *object interaction field* operates in the local coordinate frame of a specific object. The interaction field F_ϕ takes as input a simplified pose $\tilde{X}=\{\boldsymbol{j}^p, \boldsymbol{t}^p\}$, which includes joint positions and global translation. The field outputs an offset *vector* $\Delta\tilde{X}=F_\phi(\tilde{X})$ that projects the input pose to the manifold of valid interaction poses for the object such that $\tilde{X}+\Delta\tilde{X}$ is a plausible interaction pose. Fig. 3 visualizes the output vectors of an example interaction field for a chair. Querying the field with a sitting pose away from the chair (*i.e.* not a valid interaction) gives a correction pointing back towards the chair. For further away points, the visualized vectors are longer, indicating larger corrections are needed.

Guidance Objective. The object interaction field serves as a differentiable function that can be incorporated into the guidance objective to judge how far a motion is from the desired interaction manifold. Let $\tilde{X}_i \in \boldsymbol{\tau}$ be the simplified pose from the i^{th} frame of a motion $\boldsymbol{\tau}$. If we know that this pose should be a valid interaction pose, then the guidance objective is defined as $G(\boldsymbol{\tau}) = \|F_\phi(\tilde{X}_i)\|_2^2$. During denoising, output poses from the diffusion model are evaluated by this guidance objective to encourage the generated motion to contain a valid interaction pose.

Training. Supervising F_ϕ requires a dataset of invalid poses with corresponding valid interaction poses. We collect this after training the diffusion model detailed in §3.1. In particular, a noisy ground-truth interaction motion $\boldsymbol{\tau}^k$ at a random noise level k is given to the diffusion model as input. The denoiser predicts $\hat{\boldsymbol{\tau}}^0$, which *should* match the ground truth $\boldsymbol{\tau}^0$ if the model is perfect. In practice, denoising back to ground truth is difficult at high noise levels (*e.g.*, k=900), so we consider $\hat{\boldsymbol{\tau}}^0$ as an invalid interaction motion with a corresponding valid motion $\boldsymbol{\tau}^0$. Finally, we extract the last frame of the motion $\tilde{X}_N \in \hat{\boldsymbol{\tau}}^0$ as the training input for the interaction field, since this is the interaction pose in our data (see §3.3). We further augment the dataset by applying random rigid transformations to the invalid interaction poses.

Given the pose from the diffusion model $\tilde{X}_N \in \hat{\boldsymbol{\tau}}^0$ and corresponding ground truth interaction pose $\tilde{Y}_N \in \boldsymbol{\tau}^0$, the interaction field training loss is computed as $\|F_\phi(\tilde{X}_N) - (\tilde{Y}_N - \tilde{X}_N)\|_1$. Note that training on outputs from the diffusion model is important since the interaction field operates on these kinds of outputs during test-time guidance.

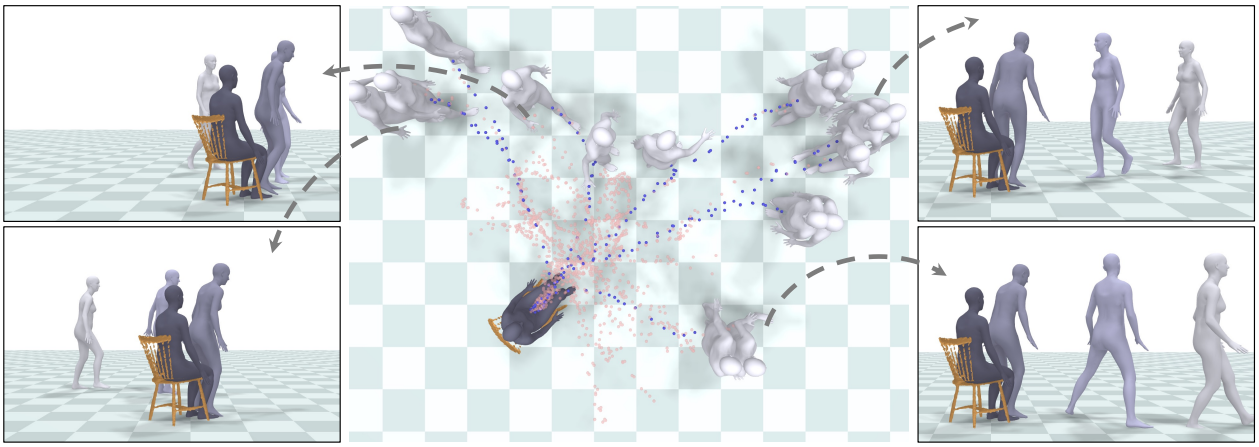

Figure 4. **Generated Synthetic Data**. Motion sequences from one tree rollout are visualized for one sitting **anchor pose**. The middle shows a bird's-eye view of the pelvis joint trajectories in **light pink**. All trajectories end in the same sitting pose, yet start at diverse locations around the chair. We highlight a few trajectories in **blue** and show full-body motions from the corresponding generations on the left and right sides. The full dataset contains many trees for different objects and humans.

Architecture. As shown in Fig. 2 (left), the interaction field architecture is an encoder-only transformer that operates on the input pose as a token. In practice, it also takes in the canonical object point cloud as a conditioning token to allow training a single field for multiple objects.

3.3. Automatic Synthetic Data Generation

We propose an automated pipeline to generate synthetic interaction motions. We first select *anchor* pose frames from an existing small dataset [3] that are indicative of an interaction we want to learn. Our key insight is to use a pre-trained scene-unaware motion model [34] to sample a diverse set of motions that *end* at a selected anchor pose, thereby demonstrating the interaction. We describe the key components next and refer to the supplement for details.

Anchor Pose Selection. We require a small set of anchor poses that each capture the key frame of an interaction motion. For sitting on a chair this is the sitting pose when the person first becomes settled in the chair (see Fig. 1). In generating motion data, these anchor poses will be the *final* frame of each synthesized motion sequence. This is because our goal is to generate motions that initiate contact with the object, but *not* to actually move the object after making contact. For the experiments in §4, these anchor frames are chosen manually from a small dataset of interactions [3].

Generating Motions in Reverse. The goal is to generate human motions that end in the chosen anchor poses and reflect realistic object interactions. We leverage HuMoR [34], which is a conditional motion VAE trained on the AMASS [27] mocap dataset. It generates realistic human motions through autoregressive rollout, but is scene-unaware. To force rollouts from HuMoR to match the final anchor pose, we could use online sampling or latent motion optimization, but these are expensive and not guaranteed to exactly converge. Instead, we re-train HuMoR as a time-reversed motion model that predicts the past instead of the future motion given a current input pose. Starting from a desired interaction anchor pose X_N, our reverse HuMoR will generate $X_{N-1}, X_{N-2}, \cdots, X_1$ forming a full interaction motion that, by construction, ends in the desired pose.

Tree-Based Rollout & Filtering. To ensure sufficient diversity and realism in motions from HuMoR, we devise a branching rollout strategy that is amenable to frequent filtering and results in a tree of plausible interactions. Starting from the anchor pose, we first sample 30 frames (1 sec) of motion. Then, multiple branches are instantiated and random rollout continues for another 30 frames on these branches independently. Continuing in this branching fashion allows growing the motion dataset exponentially while also filtering to ensure branches are sufficiently diverse and do not contain undesireable motions. Filtering involves heuristically pruning branches with motions that collide with the object, float above the ground plane, result in unnatural pose configurations, and become stationary. For our experiments, we roll out to a tree depth of 7 and sample many motion trees starting from each anchor pose. Individual paths are extracted from the tree to give interaction motions, and we post-filter out sequences that start within 1 meter of the object.

Generated Datasets. We use this scalable strategy to generate training data for sitting and lifting interactions. Fig. 4 demonstrates the diversity of our generated datasets by visualizing top-down trajectories and example motions from a single tree of sitting motions. For the sitting dataset, we choose 174 anchor pose frames across 7 subjects in the BEHAVE [3] dataset. This results in a dataset of 200K motion sequences that include sitting on chairs, stools, tables, and exercise balls. Each motion sequence in this dataset ends at a sitting anchor pose. For lifting, 72 anchor poses from 7 subjects are used to produce 110K motion sequences. Each sequence ends at a lifting anchor pose when the person ini-

tially contacts the object.

4. Experiments

NIFTY is evaluated after training on the *sitting* and *lifting* datasets introduced in §3.3. Implementation details are given in §4.1, followed by a discussion of evaluation metrics in §4.2 and baselines in §4.3. Experimental results are presented in §4.4 along with an ablation study in §4.5.

4.1. Implementation Details

We train our diffusion model M_θ for 600K iterations with a batch size of 32 using the AdamW [26] optimizer with a learning rate of 10^{-4}. A separate model is trained for sitting and lifting. We use $K{=}1000$ diffusion steps and sample the diffusion step k from a uniform distribution during training. The object interaction field F_ϕ is trained on the data described in §3.2 for 300K iterations using AdamW with a maximum learning rate of 5×10^{-5} and a one cycle LR schedule [39]. When sampling from the diffusion model, 10 samples are generated in parallel and all are guided using the object interaction field; the sample with the best guidance objective score is used as the output. We apply interaction field guidance on the last frame of motion (*i.e.* the interaction anchor pose in our datasets). Our models are trained using PyTorch [31] on NVIDIA A40 GPUs, and takes about 2 days to train on a single GPU. Visualizations use the PyRender engine [28]. At inference, gudiance with motion model takes $34s/$sample (with 10 reps).

4.2. Evaluation Setting and Metrics

To ensure we properly evaluate the generalization capability of methods trained on our synthetic interaction datasets, we *do not* create a test set using the procedure described in Sec. 3.3, which may result in a very similar distribution to training data. Instead, we create a set of 500 test scenes for each action where objects are randomly placed in the scene and the human starts from a random pose generated by HuMoR. All methods are tested on these same scenes.

Evaluating human motion coupled with object interactions is challenging and has no standardized protocol. Hence, we evaluate using a diverse set of metrics including a user perceptual study. We briefly describe the metrics next and include full details in the supplementary material.

User Study. No single metric can capture all the nuances of human-object interactions, so we employ a perceptual study [29, 42, 43, 45, 53]. For each method, we create videos from generated motions on the test scenes. To compare two methods, users are presented with two videos on the same test scene and must choose which they prefer (full user directions are in the supplement). We perform independent user studies for lifting and sitting actions using `hive.ai` [1]. Responses are collected from 5 users for every comparison video, giving 2500 total responses in each comparison study.

Foot Skating. Similar to prior work [29], we define the foot-skating score for a sequence of N timesteps as $\frac{1}{N}\sum_i^N v_i(2-2^{h_i/H}) \cdot \mathbb{1}_{h<=H}$, where v_i is the velocity and h_i is the height above ground of the right toe vertex for the i^{th} frame. H is 2.5 cm. Intuitively, this is the mean foot velocity when it is near the ground (where it *should* be 0), with higher weight applied closer to the ground.

Distance to Object (*D2O*). Similar to prior work [53], this evaluates whether the human gets close to the object during the interaction. It measures the minimum distance from the human body in the last frame of the motion sequence to any point on the object's surface. We report the % of sequences within 2 cm distance to avoid sensitivity to outliers, along with the 95^{th} percentile ($\tilde{\%}$) of this distance.

Penetration Score (% *Pen*). To evaluate realism as the human approaches an object, we measure how much penetration occurs. Based on our synthetic data, we define the first N_A frames of motion to be the approach for each action type (see supplement). Then the penetration distance for a trajectory is $\frac{1}{N_A}\sum_v \sum_i^{N_A} \texttt{sdf}_i(v) \cdot \mathbb{1}_{\texttt{sdf}_i(v)>0}$, where $\texttt{sdf}_i$ is the signed distance function of the human in the i^{th} frame and v is one of 2K points on the object's surface. We report the percentage of trajectories with penetration distance $\leq$ 2 cm (% Pen. $\leq 2cm$) ignoring trajectories with *D2O* > 2 cm, since trajectories that do not approach the object will trivially avoid penetration.

Skeleton Distance & Contact IoU. These evaluate how well generated interaction poses align with ground truth poses and their human-object contacts. We find the minimum distance between the final pose of a generated sequence and the anchor poses in the synthetic training data. The distance to this nearest neighbor pose is reported as the skeleton distance. To measure how well contacts from the generated motion match the data, we compute the IoU between contacting vertices (those that penetrate the object) on the predicted body mesh and those on the nearest neighbor mesh.

4.3. Baselines

SAMP [12]. SAMP is a strong baseline model that has been shown to produce high-quality motion for sitting on chairs after training on a large, augmented dataset for a *single character*. In the vein of other animation works [41], it is an auto-regressive VAE that consists of GoalNet to predict the final interaction position and direction on the object, and MotionNet to predict the motion of the human. We use the public training code to train a model on our challenging and diverse human-object interaction dataset. The model is trained for 4.8M iterations with scheduled sampling. Since one model is trained for each interaction type, action labels are not used in the model. At inference, we roll-out motions for up to a max-length of 300 frames and clip the motion at the frame that is closest to the goal pose for a fair comparison.

	Method	FS ↓	% *D2O* ≤ 2*cm* ↑	*D2O* 95$^{\text{th}}$% ↓	Skel. Dist. ↓	Cont. IoU ↑	% Pen. ≤ 2*cm* ↑
Sit	SAMP [12]	0.94	80.3	0.30	1.17	0.13	31.0
	cVAE [53]	0.77	88.8	0.13	1.07	0.21	46.6
	cMDM [45]	**0.36**	37.9	1.06	2.81	0.04	50.5
	NIFTY	0.47	**99.6**	**0.00**	**0.54**	**0.54**	**65.0**
Lift	SAMP [12]	0.85	59.4	0.17	1.33	0.02	38.3
	cVAE [53]	0.66	57.4	0.66	1.70	0.03	60.1
	cMDM [45]	**0.28**	36.3	1.14	2.58	0.02	46.2
	NIFTY	0.34	**77.7**	**0.05**	**0.42**	**0.17**	**68.5**

Table 1. **Quantitative Comparison.** NIFTY outperforms baselines on both sitting and lifting. Our diffusion model, guided by the learned interaction field, generates motions that reach the object (*D2O*) with few *penetrations* and realistic *contacts*. Motions are realistic with low *foot skating* and the final interaction pose is similar to synthetic data with low *skeleton distance*.

cVAE [53]. This model comes from recent work HUMANISE [53], which learns plausible human motions conditioned on scenes for four actions (lie, sit, stand, walk). This model is a conditional VAE with a GRU motion encoder and sequence-level transformer decoder. Since we evaluate on sitting and lifting actions separately, we modify their approach to remove language conditioning while keeping the scene conditioning. The model is trained on our synthetic data for 600K iterations with the recommended hyperparameters and learning rate of 10^{-4}.

cMDM [45]. This is a motion diffusion model (MDM) [45] with added object input, *i.e.*, our object-conditioned diffusion model with no interaction field guidance.

4.4. Experimental Results

User Studies. Fig. 5 shows how often users prefer our method (NIFTY) over baselines and Synthetic Data (Syn. Data) for both sitting and lifting. We perform separate studies for each comparison. Users prefer NIFTY over baselines a vast majority of the time. Averaged over both actions, NIFTY is preferred over the state-of-the-art SAMP [12] baseline 87.2% of the time and cVAE [53] 89.4% of the time, demonstrating the high quality of generated motions on a variety of objects. Similarly, NIFTY is preferred over cMDM [45] 86.3% of the time, highlighting the importance of using guidance with our interaction field during sampling. Compared to held out motions from the synthetic training data, NIFTY is preferred 47.2% of the time, which indicates that the motions are nearly indistinguishable.

To evaluate the quality of our synthetically generated sitting data, we conduct an additional study to evaluate motions from our dataset along with sitting mocap data from AMASS [27]. Users are shown a single video of a motion and asked to score the quality of the motion using a Likert scale from 1 (unrealistic) to 5 (very realistic). Motions from our synthetic dataset achieve a score of 4.39 compared to

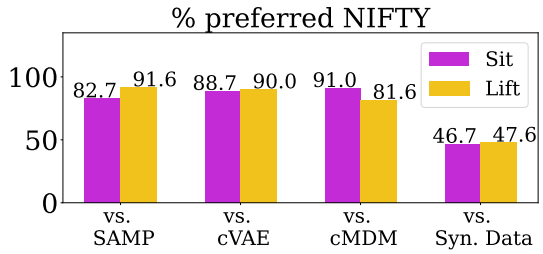

Figure 5. **User Study.** NIFTY is preferred ≥ 82.7% of the time for sitting and ≥81.6% for lifting compared to baselines. Our motions are also nearly indistinguishable from synthetic data trajectories.

4.87 for motions in AMASS, indicating the data is on par with mocap even though generated from a motion prior [34].

Quantiative Results. Tab. 1 compares NIFTY to baselines for both sitting and lifting interactions. NIFTY generates motions that reach the target object and approach realistically, as indicated by distance-to-object (*D2O*) and *penetration* metrics. Although cMDM [45] produces realistic motion with low *foot skating*, it struggles to properly approach the object since it does not use guidance from the learned interaction field. We see that interaction poses and the resulting object contacts generated by our method do reflect the synthetic dataset, resulting in low *skeleton distance* and high *contact IoU*, unlike cVAE [53] which is worse across all metrics.

NIFTY outperforms SAMP across all metrics. Motions generated by SAMP get close to objects but tend to violate object constraints when GoalNet generates goals that are not well-suited to a given starting pose, resulting in high penetration. Its autoregressive design also results in the highest foot-skating across all methods.

Qualitative Results. Fig. 6 shows a qualitative comparison between motions generated by our method and baselines. NIFTY synthesizes realistic sitting and lifting with a variety of objects. Examples show that the cMDM and cVAE baselines struggle to generalize to unseen object poses, and have no mechanism to correct for this at test time. SAMP's autoregressive nature allows some notion of feedback during generation, but still often results in object penetrations and missing contact. Our learned interactions field helps to avoid this through diffusion guidance. Please see the supplementary videos to best visualize these results.

4.5. Ablation Study

Dist. OIF. As detailed in §3.2, our object interaction field (OIF) is formulated to predict an offset vector $\Delta\tilde{X}$ that captures both distance and direction for each component of the pose state. We ablate this design decision in Tab. 2 by comparing our formulation to a version that predicts only a scalar distance to the interaction pose manifold (Distance OIF), similar to prior work [46]. As indicated by worse performance in most metrics, learning a single distance is a harder task compared to predicting an offset vector, which provides a strong learning signal for training.

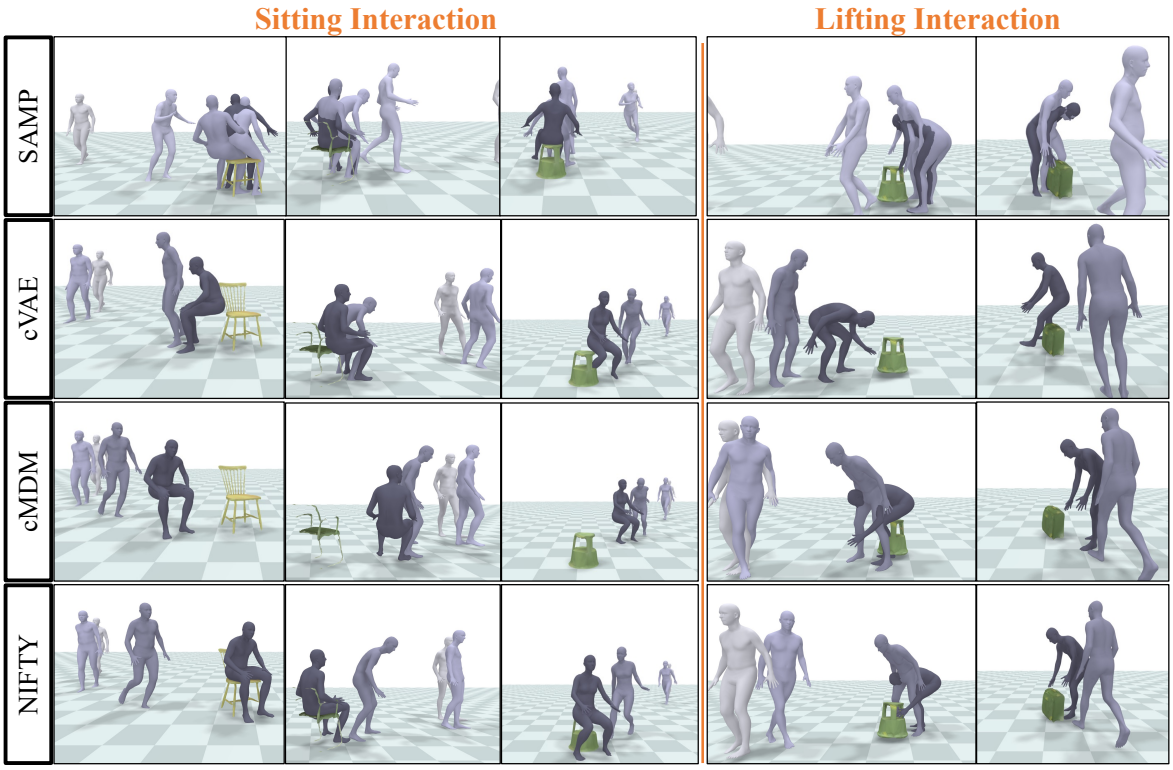

Figure 6. **Qualitative Results.** All interactions end in their own respective anchor poses and do not manipulate the object. NIFTY generates realistic interaction motions that reach the desired object with plausible contacts (*e.g.* col 1 & 4) while avoiding penetrations, unlike baselines. The mesh color gets darker as time progresses. SAMP [12] generates motion sequences that intersect with the objects(col 1,2). cVAE [53] motions have the **final interaction pose** away from the object (col 1,3,4), incorrect (col 2,5), or intersecting (col 5). cMDM [45] generates sitting poses far away from the object (col 1,3). Best viewed here.

	Method	FS ↓	% *D2O* $\leq 2cm$ ↑	*D2O* $95^{\text{th}}\%$ ↓	Skel. Dist ↓	Cont. IoU ↑	% Pen. $\leq 2cm$ ↑
Sit	Dist. OIF	**0.41**	80.9	0.47	1.25	0.24	**66.8**
	NIFTY (NN)	0.46	99.8	**0.00**	**0.28**	0.46	62.5
	NIFTY	0.47	**99.6**	**0.00**	0.54	**0.54**	65.0
Lift	Dist. OIF	**0.32**	71.1	0.07	0.52	0.11	63.3
	NIFTY (NN)	**0.32**	77.3	0.06	**0.37**	0.14	62.6
	NIFTY	0.34	**77.7**	**0.05**	0.42	**0.17**	**68.5**

Table 2. **Ablation Study.** Our full interaction field (NIFTY) predicts an offset vector is compared to an ablation that predicts a single scalar distance (Distance OIF). We also compare against a non-parameteric nearest-neighbor (NN) field that assumes access to all interaction training poses at test time.

NIFTY (NN). We also compare our interaction field to a non-parametric variant implemented using a nearest neighbor search. Specifically, during the guidance phase, we identify the nearest anchor pose from the training set and use the difference between this pose and the predicted final pose as the correction. This correction defines the distance field and guides the diffusion model accordingly. Tab. 2 shows that learning the interaction field is valuable with improved performance across most metrics including *D2O*.

5. Conclusion and Limitations

We introduced NIFTY, a framework for learning to synthesize realistic human motions involving 3D object interactions. Results demonstrate that our object-conditioned diffusion model gives improved motions over prior work when guided by a learned object interaction field and trained on automatically synthesized motion data. NIFTY opens several interesting avenues for future work. The quality of outputs depends on the pre-trained motion model [34] used in data generation; better models like [37] could improve quality. Currently, our approach does not address manipulating objects and chaining different interactions together. We leave the exploration of our interaction field to these problems to future work. Generalizing to new objects could be potentially achieved by data augmentation strategies like in [12]. Currently the model sometimes display backward walking motion and those can be fixed with better data filtering strategies. Further similar to [4] the model struggles with harder approaches (from behind the chair for instance). Moreover, we have shown results on sitting and lifting, but we would like to widen the scope to handle additional interactions by collecting new anchor poses, synthesizing data, and training our diffusion model and interaction field.

References

[1] Hive.ai. https://thehive.ai/. Accessed: 2023-11-15. 6

[2] Randall Balestriero and Yann LeCun. Police: Provably optimal linear constraint enforcement for deep neural networks. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 3

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 1, 2, 5

[4] Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5887–5895, 2021. 2, 8

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

[6] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[7] Priya L Donti, David Rolnick, and J Zico Kolter. Dc3: A learning method for optimization with hard constraints. *arXiv preprint arXiv:2104.12225*, 2021. 3

[8] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Eurographics*, 2023. 1

[9] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 3

[10] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4318–4329, 2021. 2

[11] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 1, 2

[12] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proceedings of the International Conference on Computer Vision 2021*, 2021. 1, 2, 6, 7, 8

[13] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021. 1, 2

[14] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. 2023. 2

[15] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 1, 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2, 3

[18] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 2

[19] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022. 2

[20] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[21] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022. 2, 3

[22] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9365–9376, 2023. 2

[23] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 2

[24] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Gmd: Controllable human motion synthesis via guided diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 3

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[27] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 5, 7

[28] Matthew Matl. Pyrender. https://github.com/mmatl/pyrender, 2019. 6

[29] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. *arXiv preprint arXiv:2304.02061*, 2023. 6

[30] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2, 3

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[32] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 2

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 3

[34] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5, 7, 8

[35] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. *CVPR*, 2023. 2, 3

[36] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 2

[37] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2, 8

[38] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se (3)-equivariant object representations for manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6394–6400. IEEE, 2022. 2

[39] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 6

[40] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[41] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209–1, 2019. 1, 2, 6

[42] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 1, 2, 6

[43] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13263–13273, 2022. 6

[44] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 3

[45] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *ICLR*, 2023. 1, 2, 3, 6, 7, 8

[46] Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 572–589. Springer, 2022. 2, 3, 7

[47] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. Edge: Editable dance generation from music. *arXiv preprint arXiv:2211.10658*, 2022. 2

[48] Julen Urain, Niklas Funk, Georgia Chalvatzaki, and Jan Peters. Se (3)-diffusionfields: Learning cost functions for joint grasp and motion optimization through diffusion. *arXiv preprint arXiv:2209.03855*, 2022. 2

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[50] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021. 1, 2

[51] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12206–12215, 2021. 2

[52] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022. 1, 2

[53] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 6, 7, 8

[54] Thomas Weng, David Held, Franziska Meier, and Mustafa Mukadam. Neural grasp distance fields for robot manipulation. *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 2, 3

[55] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. *arXiv:2310.08580*, 2023. 2

[56] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 1, 2

[57] Sifan Ye, Yixing Wang, Jiaman Li, Dennis Park, C Karen Liu, Huazhe Xu, and Jiajun Wu. Scene synthesis from human motion. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[58] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[59] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2

[60] Wanyue Zhang, Rishabh Dabral, Thomas Leimkühler, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. Roam: Robust and object-aware motion generation using neural pose descriptors. *arXiv preprint arXiv:2308.12969*, 2023. 2

[61] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. 2022. 1, 2

[62] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20481–20491, 2022. 2

[63] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, , and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *International conference on computer vision (ICCV)*, 2023. 2

[64] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 676–694. Springer, 2022. 2

[65] Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3560–3566. IEEE, 2023. 2