FLEX: Full-Body Grasping Without Full-Body Grasps

Purva Tendulkar Dídac Surís Carl Vondrick
Columbia University
{purvaten, didacsuris, vondrick}@cs.columbia.edu

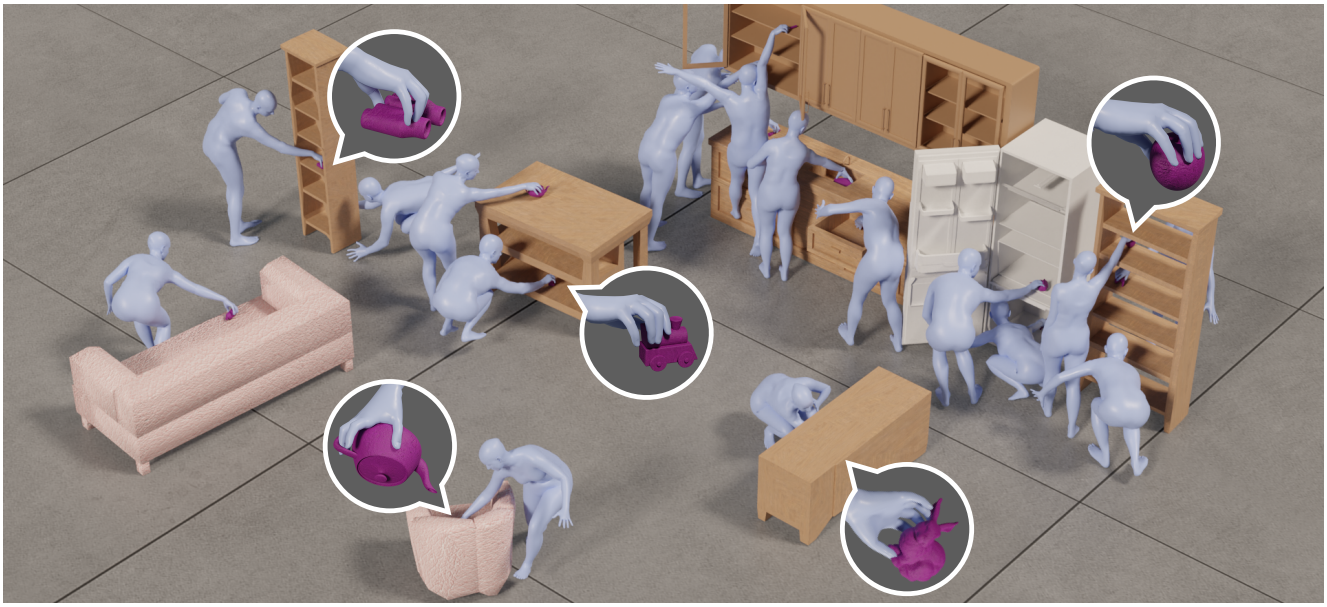


Figure 1. FLEX generates diverse full-body poses for grasping 3D objects in complex scenes. It does not use any full-body grasping data, and fully relies on body-pose priors (without grasps) and hand grasping priors (without full-body poses).

Abstract

Synthesizing 3D human avatars interacting realistically with a scene is an important problem with applications in AR/VR, video games, and robotics. Towards this goal, we address the task of generating a virtual human – hands and full body – grasping everyday objects. Existing methods approach this problem by collecting a 3D dataset of humans interacting with objects and training on this data. However, 1) these methods do not generalize to different object positions and orientations or to the presence of furniture in the scene, and 2) the diversity of their generated full-body poses is very limited. In this work, we address all the above challenges to generate realistic, diverse full-body grasps in everyday scenes without requiring any 3D full-body grasping data. Our key insight is to leverage the existence of both full-body pose and hand-grasping priors, composing them using 3D geometrical constraints to obtain full-body grasps. We empirically validate that these constraints can generate a variety of feasible human grasps that are superior to baselines both quantitatively and qualitatively. See our webpage for more details: flex.cs.columbia.edu.

1. Introduction

Generating realistic virtual humans is an exciting step towards building better animation tools, video games, immersive VR technology and more realistic simulators with human presence. Towards this goal, the research community has invested a lot of effort in collecting large-scale 3D datasets of humans [1–18]. However, the reliance on data collection will be a major bottleneck when scaling to broader scenarios, for two main reasons. First, data collection using optical marker-based motion capture (MoCap) systems is quite tedious to work with. This becomes even more complicated when objects [19] or scenes [20] are involved, requiring expertise in specialized hardware systems [21–23], as well as commercial software [24–27]. Even with the best combination of state-of-the-art solutions, this process often requires multiple skilled technicians to ensure clean data [19].

Second, it is practically impossible to capture all possible ways of interacting with the ever-changing physical world. The number of scenarios grows exponentially with every considered variable (such as human pose, object class, task,

or scene characteristics). For this reason, models trained on task-specific datasets suffer from the limitations of the data. For example, methods that are supervised on the GRAB dataset [19] for full-body grasping [28, 29] fail to grasp objects when the object position and/or orientation is changed, and generate poses with virtually no diversity. This is understandable since the GRAB dataset mostly consists of *standing* humans grasping objects at a *fixed height*, interacting with them in a relatively *small range* of physical motions. However in realistic scenarios, we expect to see objects in all sorts of configurations - lying on the floor, on the top shelf of a cupboard, inside a kitchen sink, etc.

To build human models that work in realistic scene configurations, we need to fundamentally re-think how to solve 3D tasks without needing any additional data, effectively utilizing existing data. In this paper, we address the task of generating full-body grasps for everyday objects in realistic household environments, by leveraging the success of hand grasping models [19, 30–32] and recent advances in human body pose modeling [1, 33].

Our key observation is that we can compose different 3D generative models via geometrical and anatomical constraints. Having a strong prior over full-body human poses (knowing what poses are feasible and natural), when combined with strong grasping priors, allows us to express full-body *grasping* poses. This combination leads to full-body poses which satisfy both priors resulting in natural poses that are suited for grasping objects, as well as hand grasps that human poses can easily match.

Our contributions are as follows. First, we propose `FLEX`, a framework to generate full-body grasps without full-body grasping data. Given a pre-trained hand-only grasping model as well as a pre-trained body pose prior, we search in the latent spaces of these models to generate a human mesh whose hand matches that of the hand grasp, while simultaneously handling the constraints imposed by the scene, such as avoiding obstacles. To achieve this, we introduce a novel obstacle-avoidance loss that treats the human as a connected graph which breaks when intersected by an obstacle. In addition, we show both quantitatively and qualitatively that `FLEX` allows us to generate a wide range of natural grasping poses for a variety of scenarios, greatly improving on previous approaches. Finally, we introduce the ReplicaGrasp dataset, built by spawning 3D objects inside ReplicaCAD [34] scenes using Habitat [35].

2. Related Work

Object grasping. Generating a 3D hand for grasping objects is a widely studied task in robotics [31, 36–43], graphics [44–51] and 3D computer vision [19, 30, 52–55]. Many works have also studied the anatomy of human hands to create grasp taxonomies [33, 56–59]. Most existing work tries to imitate static hand-grasping positions from

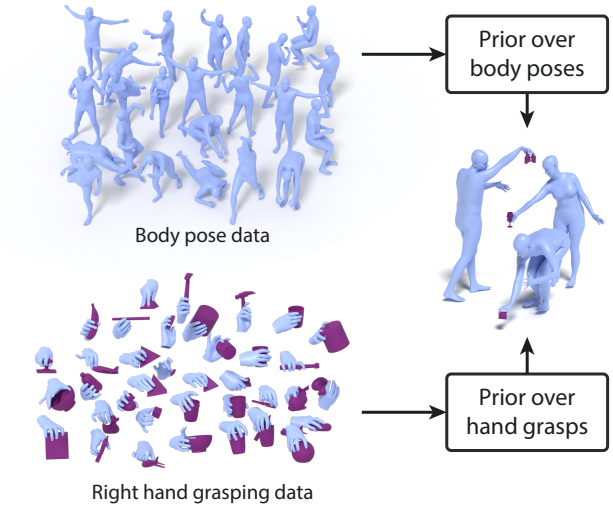


Figure 2. **Overview**. We leverage existing body pose priors and hand-grasping models (left) to perform full-body grasping in complex scenes (right). Our method does not rely on any data for full-body grasping and surpasses methods requiring it, in terms of diversity and generalization to complex scenes.

data [19, 30, 52, 53], while other efforts generate stable dynamic grasps using either motion-based grasping data [32] or reinforcement learning [60]. Recently, Turpin *et al.* [31] used a differentiable physics simulator to generate grasps that are physically stable. All these methods have looked at generating grasps for objects in presence of simple or no obstacles (*e.g.*, placed on a counter-top). Instead, we focus on generating hand grasps for objects in more realistic scenarios (*e.g.*, objects in refrigerator, drawers, kitchen sink).

Full-body grasping. Instead of synthesizing just hand-object interactions, the community has recently started generating full-body interactions with objects. GRAB dataset [19] for full-body object interaction was collected using motion capture. GOAL [28] and SAGA [29] use GRAB to build generative models for full-body grasping. Synthesizing full-body interactions has also been explored to create simulators (VirtualHome) for studying house-hold activities [61]. VirtualHome uses pre-defined animations but they are not very realistic and intersect with other objects. In this work, we generate realistic full-body grasps *without using full-body data and in presence of obstacles.*

Test-time optimization. Test-time optimization has been used to improve neural network generalization by either using task-specific constraints or self-supervision during inference [62–67]. This principle is applied to the hand-object grasping problem [19, 28–30] by first generating a coarse grasp and later refining it using test-time optimization. These methods are especially suitable for tasks where constraints are easy to specify. Zhang *et al.* [68] use "3D common sense" constraints to resolve ambiguity in 3D spa-

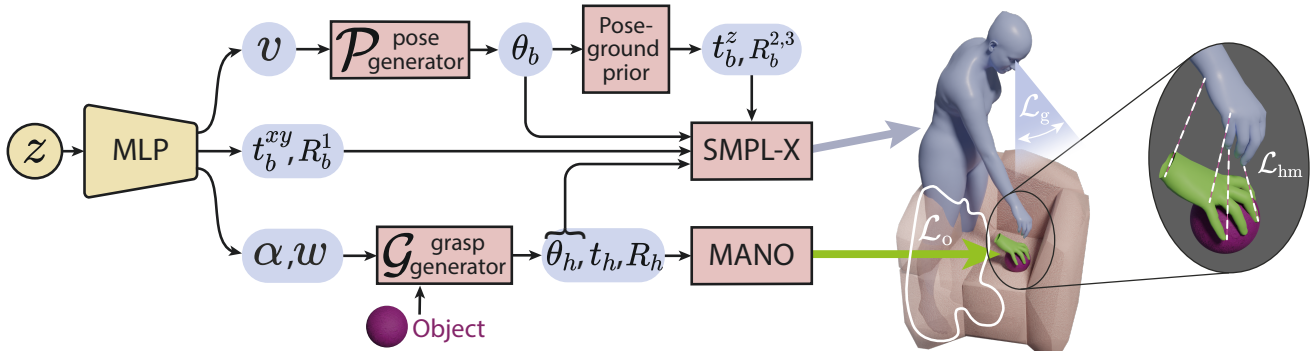


Figure 3. **Method**. Given pre-trained hand-grasping and human pose priors, our method performs a gradient-based search procedure over five different landscapes to minimize the hand matching, obstacle and gaze losses. Additionally, our data-driven pose-ground prior ensures that the pose is reasonable with respect to the ground. In the figure, the parameters we optimize are shown in yellow, the differentiable (frozen) layers are shown in pink, and the activations are shown in blue.

tial arrangements of humans and objects jointly from 2D, while Liu *et al.* [69] use shadows as a constraint for inferring the 3D structure of an occluded object. Similarly, we obtain full-body grasps by combining different priors satisfying intuitive geometrical constraints.

3. ReplicaGrasp Dataset

Existing benchmarks for full-body object grasping [19] only consider objects in a limited height range without any other objects in their vicinity. To make the task more challenging and representative of the real-world, we build the ReplicaGrasp dataset. ReplicaGrasp contains 50 everyday objects from GRAB (such as wineglass or cellphone) present in 48 receptacles of ReplicaCAD [34], simulated with the Habitat simulator [35] to be in a variety of feasible positions, leading to a total of 4.8k instances. The receptacles include surfaces of both rigid and articulated furniture items, such as drawers, which may be open to different degrees, leading to interesting cases like objects being deep inside on the bottom shelf of refrigerator, or on the top shelf of a cupboard. We refer to the furniture on which the target object to-be-grasped is spawned as the 'obstacle' for that instance. Succeeding on this dataset requires generating full-body human grasps that are "scene-aware"—not intersecting with obstacles—, as well as natural and feasible.

4. Approach

Given a 3D object mesh and a set of 3D obstacle meshes in its vicinity, our goal is to generate a 3D human mesh grasping the object without intersecting with the obstacles.

4.1. Preliminaries

Hand model. We use the MANO [70] differentiable 3D hand model, that takes as input the full-finger articulated pose $\theta_\text{h} \in \mathbb{R}^{15\text{x}3}$, the wrist translation $t_\text{h} \in \mathbb{R}^3$, and the wrist global orientation $R_\text{h} \in \mathbb{R}^3$, and outputs a 3D mesh $\mathcal{M}_\text{h}$, with vertices $\mathcal{V}_\text{h}$ in a global coordinate system.

Human body model. We use the SMPL-X [33] statistical 3D whole-body model, which jointly represents the body, head, face and hands. SMPL-X is a differentiable function that takes as input the full-body pose $\theta_\text{b} \in \mathbb{R}^{21\text{x}3}$, the full-finger articulated pose $\theta_\text{h} \in \mathbb{R}^{15\text{x}3}$, the pelvis translation $t_\text{b} \in \mathbb{R}^3$ and orientation $R_\text{b} \in \mathbb{R}^3$, and optionally, body shape parameters and facial expression, and outputs a 3D mesh $\mathcal{M}_\text{b}$, with vertices $\mathcal{V}_\text{b}$ in a global coordinate system.

Pre-trained generative models. We use a pre-trained generative model $\mathcal{P}$ that has learned a prior over body poses. It takes as input a latent vector v and generates a body pose which can be used as input to SMPL-X. Similarly, $\mathcal{G}$ generates right hand grasps. It takes as input a latent vector w, an approaching angle α and an object $\mathcal{O}$, represented by its vertices, and generates a hand pose, including its translation and rotation, to be used as input to MANO. We implement $\mathcal{P}$ using VPoser [33] and $\mathcal{G}$ using GrabNet [19].

4.2. Method

Given a pre-trained right hand grasping model $\mathcal{G}$ that can predict global MANO parameters $\{\theta_\text{h}, t_\text{h}, R_\text{h}\}$ for a given object, as well as a pre-trained pose prior $\mathcal{P}$ that can generate feasible full-body poses θ_b, our approach called `FLEX` (**F**ull-body **L**atent **Ex**ploration) generates a 3D human grasping the desired object. To do so, `FLEX` searches in the latent spaces of $\mathcal{G}$ and $\mathcal{P}$ to find the latent variables w and v, respectively, as well as over the space of approaching angles α, SMPL-X translations t_b and global orientations R_b, which are represented in 'yaw-pitch-roll' format. See Fig. 3 for an overview of the method.

This search, or latent-space exploration, is done via model inversion, by backpropagating the gradient of a loss at the output of our model, and finding the inputs that minimize it. This procedure is done iteratively, until the loss is minimized. During the search, we keep the weights of $\mathcal{G}$ and $\mathcal{P}$ frozen. Therefore, we do not perform any training; the procedure described in this section takes place at inference time. We describe the losses we use next.

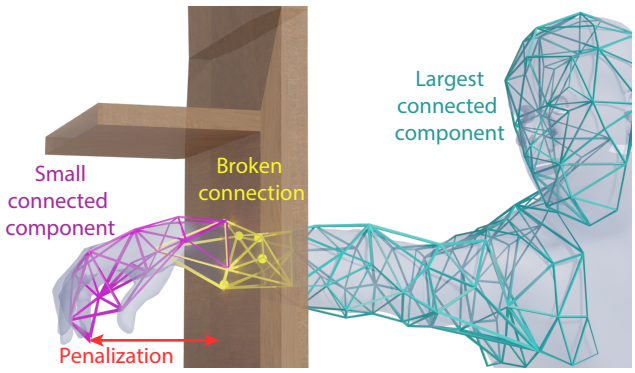


Figure 4. **Body connectivity**. When a body crosses an obstacle, the obstacle divides the body in two or more components, the break happening at the vertices that lie inside of the obstacle (in yellow). We penalize the vertices in all the connected components of the resulting body graph other than the largest one.

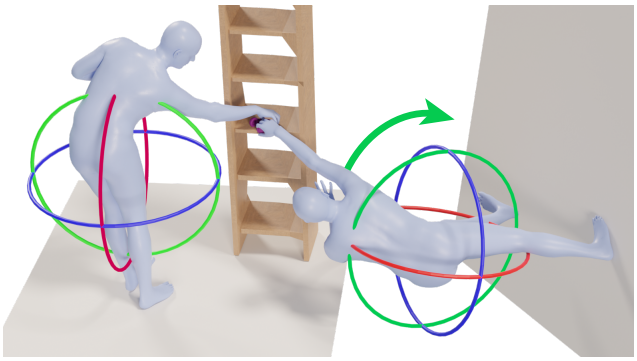


Figure 5. **Pose-ground prior**. Given a body pose, the position of the ground can be predicted. This determines the pitch and the roll of the body orientation, removing two degrees of freedom from our optimization. In this example, the human on the right should rotate to their right (pitch) and slightly forwards (roll) to get to a correct orientation with respect to the ground.

4.3. Losses

Hand matching loss. The main intuition of our paper is that we can combine generative models using constraints based on the geometry of their outputs. Specifically, we can combine the generated hand and body meshes because there exists a connection between them: the hands have to match. Given vertices $\mathcal{V}_\text{b}$ (output from SMPL-X) and $\mathcal{V}_\text{h}$ (output from MANO), we align them by minimizing:

$$\mathcal{L}_\text{hm} = \frac{1}{|\mathcal{V}_\text{h}|} \sum_{i=1}^{|\mathcal{V}_\text{h}|} d_{vv}(\mathcal{V}_{h_i}, \mathcal{V}_{b_i}^h), \tag{1}$$

where $d_{vv}(.,.)$ is the L^2 distance between two vertices in the 3D space. $\mathcal{V}_b^h$ represents the vertices within $\mathcal{V}_b$ that correspond to the right hand—the rest are not used in this loss.

Obstacle loss. To succeed at grasping objects in a scene, the humans that are generated need to avoid all the obstacles. We introduce a novel obstacle-avoidance loss that penalizes body-obstacle penetration, and consists of two parts: body-obstacle intersection and body mesh connectivity.

1. Body-obstacle intersection loss. We represent obstacles as watertight meshes, and compute the signed distance from every vertex in $\mathcal{V}_\text{b}$ to the obstacle mesh: negative distances represent points *inside* the obstacle. This loss sums the absolute values of the vertex-obstacle distances for all the vertices that lie inside the obstacle:

$$\mathcal{L}_\text{o}^\text{inter} = \frac{1}{|\mathcal{V}_\text{b}|} \sum_{i=1}^{|\mathcal{V}_\text{b}|} \left| \min\left(0, d_\text{vm}(\mathcal{V}_{\text{b}_i}, \mathcal{M}_\text{obstacle})\right) \right|, \tag{2}$$

where $d_\text{vm}(.,.)$ is the signed distance function between a vertex (3D point) and a mesh.

2. Body mesh connectivity loss. The previous loss alone does not penalize the parts of the body that penetrate the obstacle and resurface at the other side of the obstacle since these vertices are considered as "outside" vertices. To penalize them, we treat the human mesh as a graph. When part of the body penetrates an obstacle, two (or more) components of this graph get separated by the obstacle, resulting in a disconnected graph, composed of multiple connected components. The graph breaks precisely at the vertices that are inside of the obstacle (see Fig. 4). Therefore, we penalize all the vertices that are not part of the largest connected components of the resulting graph, and the penalization is the distance from those vertices to the obstacle. Mathematically, the body mesh connectivity loss is:

$$\mathcal{L}_\text{o}^\text{con} = \frac{1}{|\mathcal{V}_\text{b}|} \sum_{i=1}^{|\mathcal{V}_\text{b}^\text{discon}|} d_\text{vm}(\mathcal{V}_{\text{b}_i}^\text{discon}, \mathcal{M}_\text{obstacle}), \tag{3}$$

where the sum is over the set of vertices $\mathcal{V}_\text{b}^\text{discon} \subset \mathcal{V}_\text{b}$ that are outside of the obstacle mesh and belong to connected components of the resulting graph that are not the largest connected component, such as the pink one in Fig. 4. For efficiency, this loss uses a subsampled version of $\mathcal{M}_\text{b}$. The final obstacle loss is the sum of the two: $\mathcal{L}_\text{o} = \mathcal{L}_\text{o}^\text{inter} + \mathcal{L}_\text{o}^\text{con}$.

Gaze loss. In addition to the main losses described above, we also incorporate an explicit gaze loss to encourage the human to look at the target object. We use the head direction vector from Taheri *et al.* [28] which goes from the back to the front of the head $V_{b\rightarrow f}$. Then we compute the vector from the back of the head to the target object $V_{b\rightarrow o}$. The gaze loss minimizes the angle between these two vectors:

$$\mathcal{L}_\text{g} = \cos^{-1} \frac{V_{b\rightarrow f} \cdot V_{b\rightarrow o}}{|V_{b\rightarrow f}||V_{b\rightarrow o}|} \tag{4}$$

During inference, we minimize a combination of the three losses: $\mathcal{L} = \lambda_\text{hm}\mathcal{L}_\text{hm} + \lambda_\text{o}\mathcal{L}_\text{o} + \lambda_\text{g}\mathcal{L}_\text{g}$.

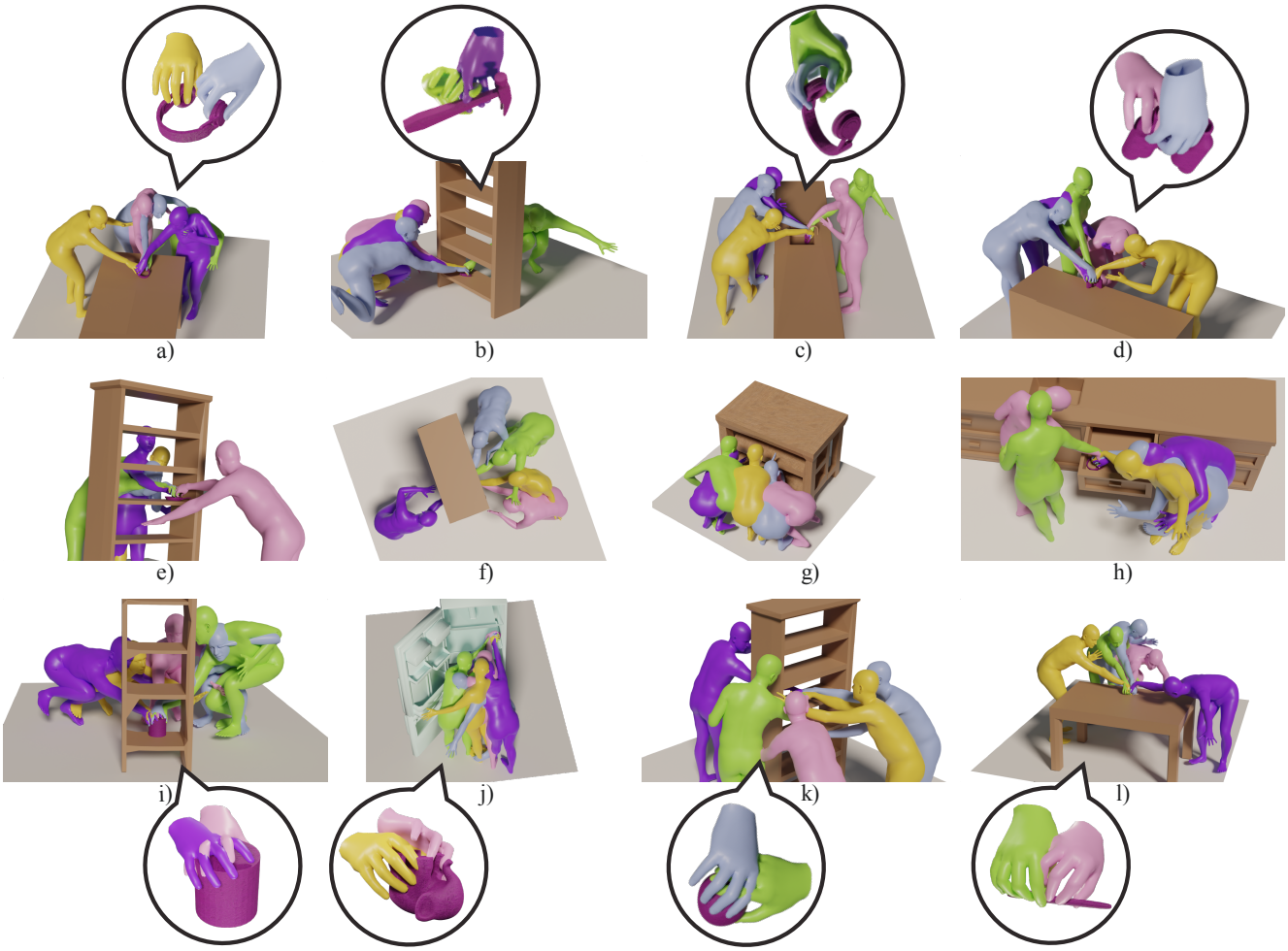


Figure 6. **FLEXibility**. We showcase FLEX's ability to generate a variety of diverse, feasible full-body grasps in a number of challenging scenarios. Different colors indicate the top-5 generations. Bubbles zoom in on diverse hand grasps.

4.4. Pose-Ground Prior

Enforcing the previous losses can lead to perfect grasps, but potentially "flying" humans. To ensure that the human pose is also reasonable with respect to the ground, we learn a prior over the relationship between human pose and ground location. Specifically, we use the AMASS dataset [1,71], which uses the *xy*-plane as the floor, to train a 2-layer MLP that predicts the roll and pitch components of the human pelvis orientation with respect to the floor, given a human pose θ_{b}. This MLP is trained using MSE regression, and it is kept frozen during inference. As exemplified in Fig. 5, given a specific human pose, the only actual degree of freedom for the human to have a natural orientation with respect to the ground is the yaw (shown in blue in the figure). The pitch and roll are constrained by the position of the ground.

Additionally, we fix the vertical component of the human body translation t_{b}^z such that the lowest vertex of the predicted body mesh touches the ground.

4.5. One Latent to Rule Them All

Our framework requires optimizing the values of several interdependent parameters—for example, the angle of the hand grasp α constrains the human body orientation R_{b} and pose v. Making sense of these relationships is required in order to smoothly search over the different parameters. Therefore, we delegate the burden of controlling the multiple parameters to a single controllable latent vector z that can abstract away the low-level dependencies of the individual latent variables. Specifically, following Liu *et al.* [72], we learn a mapping network (MLP in Fig. 3) from the latent vector z to the different parameters and latents we defined above. At inference time, we optimize the values of z and the weights of the MLP. The rest of the parameters are given as activations (outputs) of the MLP.

For every example (scene and object), we optimize N latent vectors (z), with different initializations, and at the end of the optimization process, we select the ones that result in the smallest loss. The parameters of the mapping network (a 2-layer MLP) are shared across the N latent vectors. See Appendix A.2 for more implementation details.

Method	ReplicaGrasp						GRAB				
	Obj Cont (%) ↑	Obj Penet (%) ↓	Obst Penet (%) ↓	Div_{samp} (cm) ↑	Div_{all} (cm) ↑	Ground (cm) ↓	Obj Cont (%) ↑	Obj Penet (%) ↓	Div_{samp} (cm) ↑	Div_{all} (cm) ↑	Ground (cm) ↓
GOAL [28]	1.14	2.14	6.87	0.20	37.89	5.31	1.62	3.50	0.11	7.87	3.56
SAGA [29]	1.21	**0.29**	7.27	1.10	43.84	9.40	**2.19**	3.43	1.06	17.35	2.39
FLEX (ours)	**2.20**	2.49	**0.53**	**10.37**	**69.78**	**0.00**	1.63	**2.73**	**24.94**	**46.46**	**0.00**

Table 1. **Results for ReplicaGrasp and GRAB**. We report contact and penetration with the object, obstacle penetration, diversity and distance from ground. FLEX almost always outperforms baselines, even though it was not trained on fully-body grasp data (GRAB).

5. Experiments

We conduct experiments on two datasets: our challenging 1) ReplicaGrasp and 2) GRAB [19]. GRAB is a MoCap dataset of humans interacting with everyday objects without any other obstacles. Since baselines are trained and evaluated on GRAB, we evaluate if FLEX can perform comparably despite it not using GRAB's full-body grasping data.

5.1. Baselines

We compare with the only two existing works that perform full-body human grasping: GOAL [28] and SAGA [29]. Both methods use a conditional variational auto-encoder (cVAE) [73] to reconstruct 3D humans conditioned on the object's position and orientation, and have been trained on the full-body grasps of GRAB. GOAL reconstructs a sub-sampled set of body vertices, while SAGA reconstructs surface body markers. Both use test-time optimization for refining the full-body grasp by leveraging fine-grained human–object contact information. Since these methods do not work when the objects are out of distribution in the horizontal xy-plane, we evaluate them for objects placed at $(x, y) = (0, 0)$. We then translate the resulting human-object pair to the correct (x, y) for visualization.

5.2. Ablations

To understand the impact of the key elements of our approach, we perform ablation studies on a validation set of ReplicaGrasp containing 500 random object configurations. We freeze the main optimization parameters independently: 1) hand-grasp latent θ_{h}, and 2) human pose latent θ_{b}. Additionally, we perform ablations to study the effect of: 3) removing the obstacle loss, and 4) not enforcing a pose-ground prior.

5.3. Metrics

Similar to prior work [28] we report object-grasping metrics, and additionally report obstacle and ground metrics:

- **Object contact percentage** - percentage of object vertices in contact with the human mesh,
- **Object penetration percentage** - percentage of object vertices penetrated by the human mesh.
- **Obstacle penetration percentage** - percentage of human vertices penetrating the obstacle mesh.
- **Ground distance** - absolute vertical distance from the lowermost vertex of the human mesh to the ground plane.

Higher object contact implies a more stable grasp, while lower obstacle and object penetration is naturally more preferred. We only compute object penetration for instances with non-zero contact. Note that these metrics are not perfect – e.g., lower contact could very well lead to a stable grasp. However, these are reasonably accepted proxies for automatically evaluating grasps [19, 28–30].

To evaluate the ability of different methods to generate diverse full-body grasps, following [29] we also report:

- **Sample diversity** - average vertex-to-vertex L2 distance for each sampled pair of human meshes for a single instance, averaged across instances.
- **Overall diversity** - average pairwise diversity across all pairs of generated samples for all instances of the dataset, which quantifies the method's ability to generate a range of complex human poses.

6. Results

6.1. Comparison to Baselines

The main results comparing our method FLEX to the two baselines (GOAL [28] and SAGA [29]) are shown in Tab. 1.

FLEX generates diverse full-body grasps, even with obstacles. On ReplicaGrasp, which tests grasping objects in more realistic scenarios, FLEX achieves the best object contact score. This is because in many cases where the object is either too high or too low, the baselines often fail to even touch the object. We observed that SAGA performs best on object penetration. This is explained by an explicit optimization to minimize inference-time collision loss for *1500 iterations*. However, this procedure may lead to unnatural humans – for example, humans that penetrate the ground to grasp lower objects (see Fig. 8 i,k) or elongated humans for grasping higher objects (see Fig. 8 a). This phenomenon is reflected in the ground distance metric, where SAGA generates humans on average 9 cm away from the ground. Instead, FLEX, which combines the full-body, hand-grasping, and pose-ground priors, all of which are data-driven, will be in-distribution with the data by construction. Regarding diversity, FLEX outperforms baselines by a significant margin, while also avoiding obstacles.

#	Method	Obj Cont (%) ↑	Obst Penet (%) ↓	Ground (cm) ↓	Div_{samp} (cm) ↑	Div_{all} (cm) ↑
1	Ours	2.19	0.97	0.00	10.15	70.44
2	No grasp optimization	0.88	0.85	0.00	13.19	71.07
3	No pose optimization	2.03	1.87	0.00	0.27	41.55
4	No obstacle loss	2.38	10.57	0.00	21.60	71.86
5	No pose-ground prior	2.13	0.61	71.72	16.88	96.27

Table 2. **Ablation Studies of different components FLEX**. For each metric, red represents a significant drop.

GRAB. On the GRAB dataset, FLEX performs comparably to both baselines on the object contact and penetration metrics. Note that FLEX achieves similar performance (surpassing GRAB on both metrics, and SAGA on one) *without using any full-body grasping data*. Additionally, FLEX surpasses both baselines by large margins on diversity metrics.

6.2. Ablations

We show the importance of different components of FLEX by demonstrating how performance is negatively impacted when each is removed in Tab. 2.

– No hand grasp optimization leads to the worst hand grasp with minimal contact. When we do not allow search in the latent space w of the hand-grasping model (Row 2), we get a poor grasp. Freezing the hand grasp severely restricts the space of feasible poses, leading to a deadlock between the hand-matching and obstacle losses. Hence, avoiding obstacles leads to the object being touched minimally. We further demonstrate the importance of searching for hand-grasps in Fig. 7. While all the hand-grasps are plausible, upon hallucinating the full-body human accompanying the grasp, it is evident that the human will penetrate the scene for the blue and orange grasps. Thus, searching over hand grasps will allow the model to pick grasps that lead to minimal object penetration.

– No body pose optimization leads to high obstacle penetration and lower diversity. By not searching over the full body latent space v, the model is incapable of *FLEXibly* adjusting the human pose to avoid obstacles and thus generates humans that intersect with them. For the same reason, there is little variation in overall full-body grasp predictions as evident from the low diversity numbers.

– No obstacle loss leads to high obstacle penetration. Comparing Row 4 with Row 1, we see that removing the obstacle loss (Eqs. (2) and (3)) leads to very high obstacle penetration. Because humans are not incentivized to avoid obstacles, there are more feasible body poses, which explains the significant increase in diversity.

– Removing pose-ground prior results in flying humans. Removing the pose-ground prior (Row 5) no longer forces humans to touch the ground, resulting in a very large distances (71.7 cm) from the ground.

Figure 7. **Hand grasp search.** We illustrate the benefit of searching in the latent space of the hand grasping model. Hallucinating the full body associated with each grasp makes the choice easier.

6.3. Qualitative Results

Diversity. Fig. 6 showcases FLEX's ability to discover a variety of reasonable grasps while simultaneously satisfying obstacle constraints. For example, in Fig. 6 d,l), FLEX generates both crouched and standing humans in different positions and orientations without penetrating the obstacle or compromising on the quality of the right hand grasp. When the object is very low, as in Fig. 6 b), FLEX generates squatting as well as kneeling positions.

Comparison to baselines. Qualitative results for FLEX, SAGA and GOAL are shown in Fig. 8 in yellow, orange and blue colored humans, respectively. GOAL's grasps lack diversity – the right hand always grasps the object from the top with very little variation in the distance to the object as well as the overall pose. When the object is too low, it starts to generate unnatural legs (see Fig. 8 i,j,k).

SAGA generates more diverse hand grasps (side and bottom grasps in Fig. 8 c,d,h) and full-body poses (slightly bent legs in Fig. 8 j). However, it often fails when objects are too low or too high (see Fig. 8 a,i,k).

FLEX is able to generate suitable poses based on the object placement with respect to the scene. For example, when the object is deep inside the refrigerator (see Fig. 8 f,i), FLEX can generate humans with a carefully outstretched hand such that it doesn't collide with any obstacles. In Fig. 8 a), where the object is very high on the shelf, FLEX generates a human on their toes while in Fig. 8 c), FLEX generates a good right hand grasp that carefully avoids the protruding component in the top compartment of the fridge. Fig. 8 e) demonstrates how FLEX can match the hand grasp of the baselines, and additionally jut out the rest of the body to avoid the obstacle when necessary.

Failures. Fig. 10 shows some failures. Most of them are caused by the limitations of the pre-trained generative models $\mathcal{G}$ and $\mathcal{P}$. Fortunately, FLEX is model agnostic, so it will only improve as better generative models are developed. FLEX only requires them to be differentiable.

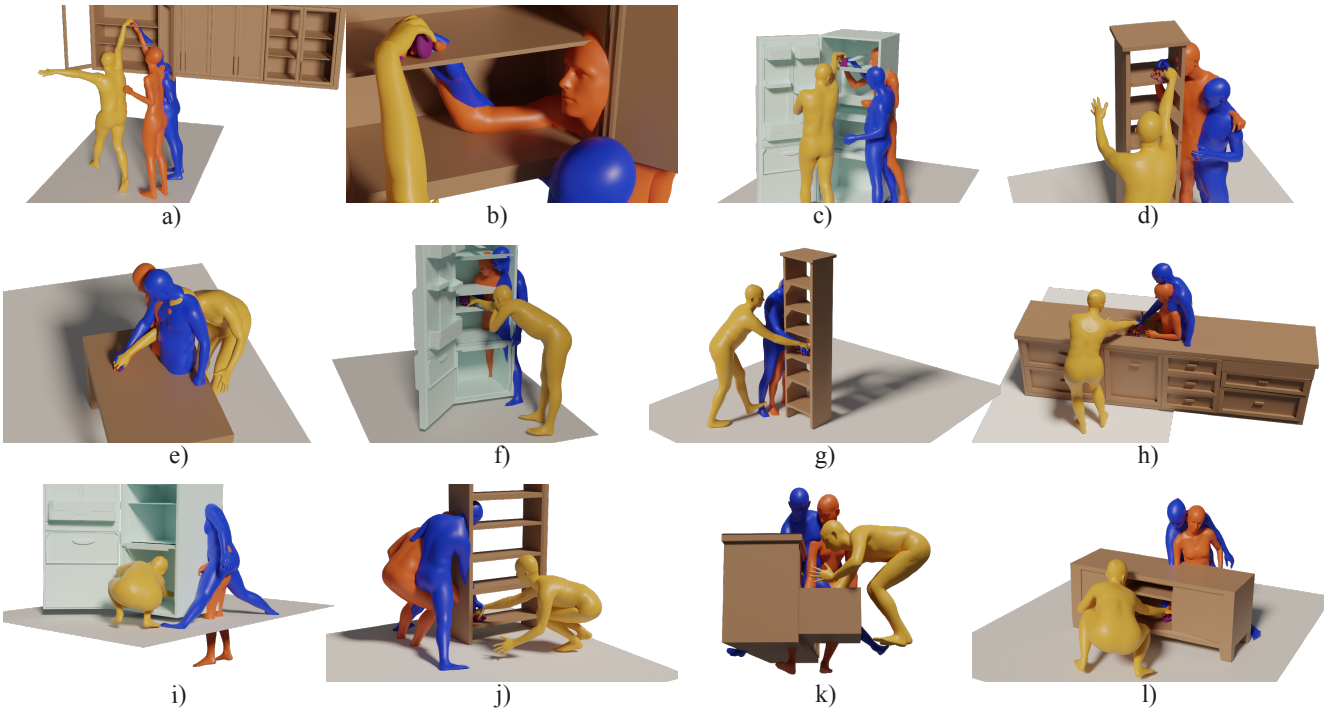


Figure 8. **Qualitative Comparisons**. We show 3D human avatars generated by FLEX, SAGA and GOAL when objects in the ReplicaGrasp dataset are placed at high **(top)**, medium **(middle)** and low **(bottom)** heights. FLEX generates a variety of reasonable poses for grasping the target objects in a number of scenarios, while SAGA and GOAL fail to generalize especially at extreme heights.

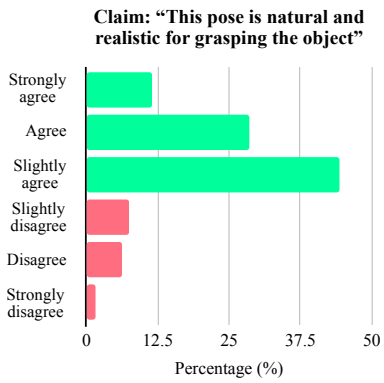


Figure 9. **Human Studies** We conduct human studies on Amazon Mechanical Turk and ask subjects to rate 3D human grasps generated with FLEX on a Likert scale of 1-6 for realism. Subjects agree 85% of the time that our generated humans grasp objects realistically.

6.4. Human Evaluation

To evaluate the perceptual quality of our generated full-body grasps, we conducted human studies on Amazon Mechanical Turk (AMT). We chose a subset of FLEX results covering all 48 receptacles twice for objects in both fallen and upright orientations. We showed each result to five different subjects on AMT in an interactive 3D interface and asked them to rate the full-body grasps on a Likert scale of 1 (strongly disagree) to 6 (strongly agree). We filtered out responses with standard deviation > 1. Results are shown in Fig. 9. 85% of the time participants agree that our generations are natural and realistic.

A subject who gave a rating of 2 wrote: *"[...] arm that is grabbing is reaching way too hard to reach object, but the grasp is natural"*, while a subject who gave a rating of 6 wrote: *"The bend on the spine and the legs being approximate shoulder width apart is a natural body movement"*.

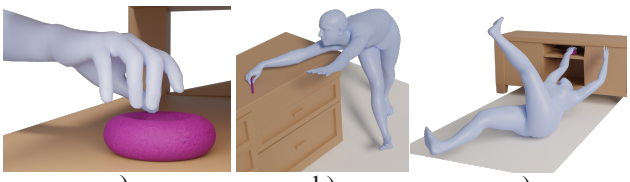


Figure 10. **Failure cases**. **a)** Our selected grasping model $\mathcal{G}$ fails. **b)** All our constraints are satisfied, but the human pose is too stretched / unnatural. **c)** The pose-ground prior may be imperfect or result in rare poses while satisfying other constraints.

7. Conclusion

In this work, we address the task of generating full-body humans grasping 3D objects in the presence of obstacles and introduce a new dataset, ReplicaGrasp, to evaluate the realism of the generations. We describe an optimization-based approach that leverages existing hand-grasping models and human pose priors to solve this task, without using any 3D full-body grasping data. Experiments show that we are able to generate realistic avatars that surpass existing methods, both quantitatively and qualitatively.

Acknowledgements: This research is based on work partially supported by NSF NRI Award #2132519, and the DARPA MCS program under Federal Agreement No. N660011924032. D.S. is supported by the Microsoft PhD fellowship. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

References

[1] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019. 1, 2, 5

[2] Advanced Computing Center for the Arts and Design. ACCAD MoCap Dataset. 1

[3] Andreas Aristidou, Ariel Shamir, and Yiorgos Chrysanthou. Digital dance ethnography: Organizing large dance collections. *J. Comput. Cult. Herit.*, 12(4), November 2019. 1

[4] Bio Motion Lab. BMLhandball Motion Capture Database. 1

[5] Nikolaus F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5):2–2, September 2002. 1

[6] Carnegie Mellon University. CMU MoCap Dataset. 1, 12

[7] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5573–5582, July 2017. 1

[8] Eyes JAPAN Co. Ltd. Eyes Japan MoCap Dataset. 1

[9] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. MoVi: A large multipurpose motion and video dataset, 2020. 1

[10] Anargyros Chatzitofis, Leonidas Saroglou, Prodromos Boutis, Petros Drakoulis, Nikolaos Zioulis, Shishir Subramanyam, Bart Kevelham, Caecilia Charbonnier, Pablo Cesar, Dimitrios Zarpalas, et al. Human4d: A human-centric multimodal dataset for motions and immersive media. *IEEE Access*, 8:176241–176262, 2020. 1

[11] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(4):4–27, March 2010. 1

[12] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour. The KIT whole-body human motion database. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 329–336, July 2015. 1

[13] Matthew Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and Shape Capture from Sparse Markers. *ACM Trans. Graph.*, 33(6), November 2014. 1, 12

[14] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database HDM05. Technical Report CG-2007-2, Universität Bonn, June 2007. 1

[15] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455, June 2015. 1, 12

[16] Simon Fraser University and National University of Singapore. SFU Motion Capture Database. 1

[17] Ludovic Hoyet, Kenneth Ryall, Rachel McDonnell, and Carol O'Sullivan. Sleight of hand: Perception of finger motion from reduced marker sets. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D '12, page 79–86, New York, NY, USA, 2012. 1

[18] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 14.1–14.13, September 2017. 1

[19] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 6, 12

[20] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, pages 2282–2292, October 2019. 1

[21] Structure sensor: 3d scanning, augmented reality and more. 1

[22] Kinect for xbox one. 1

[23] Vicon vantage: Cutting edge, flagship camera with intelligent feedback and resolution. 1

[24] Vicon shogun: Vfx motion capture. 1

[25] Poser: 3d rendering and animation software. 1

[26] Skanect: 3d scanning. https://skanect. 1

[27] Monocle: Kinect data capture app. 1

[28] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4, 6

[29] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 6

[30] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, 2021. 2, 6

[31] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *European Conference on Computer Vision*, pages 201–221. Springer, 2022. 2

[32] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision (ECCV)*. Springer, October 2022. 2

[33] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3, 12

[34] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34:251–266, 2021. 2, 3, 12

[35] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 12

[36] M.R. Cutkosky. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Transactions on Robotics and Automation*, 1989. 2

[37] Renaud Detry, Dirk Kraft, Anders Glent Buch, Norbert Krüger, and Justus H. Piater. Refining grasp affordance models by experience. *2010 IEEE International Conference on Robotics and Automation*, pages 2287–2293, 2010. 2

[38] Kaijen Hsiao and Tomas Lozano-Perez. Imitation learning of whole-body grasps. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006. 2

[39] Robert Krug, Dimitar Dimitrov, Krzysztof Charusta, and Boyko Iliev. On the efficient computation of independent contact regions for force closure grasps. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010. 2

[40] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Generating grasp poses for a high-dof gripper using neural networks. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1518–1525. IEEE, 2019. 2

[41] K. Bernardin, K. Ogawara, K. Ikeuchi, and R. Dillmann. A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. *IEEE Transactions on Robotics*, 2005. 2

[42] Andrew T. Miller and Peter K. Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11:110–122, 2004. 2

[43] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019. 2

[44] Maciej Kalisiak and Michiel Panne. A grasp-based motion planning algorithm for character animation. *The Journal of Visualization and Computer Animation*, 2001. 2

[45] Paul G. Kry and Dinesh K. Pai. Interaction capture and synthesis. *ACM Trans. Graph.*, 2006. 2

[46] Ying Li, Jiaxin L. Fu, and Nancy S. Pollard. Data-driven grasp synthesis using shape matching and task-based pruning. *IEEE Transactions on Visualization and Computer Graphics*, 2007. 2

[47] Nancy S. Pollard and Victor B. Zordan. Physically based grasping control from example. In *Symposium on Computer Animation*. The Eurographics Association, 2005. 2

[48] Hans Rijpkema and Michael Girard. Computer animation of knowledge-based human grasping. *SIGGRAPH Comput. Graph.*, 1991. 2

[49] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Trans. Graph.*, 2021. 2

[50] Oliver Glauser, Shihao Wu, Daniele Panozzo, Otmar Hilliges, and Olga Sorkine-Hornung. Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Trans. Graph.*, 2019. 2

[51] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758), 2019. 2

[52] 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

[53] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021. 2

[54] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision*, pages 361–378. Springer, 2020. 2

[55] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019. 2

[56] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on Human-Machine Systems*, 2016. 2

[57] Noriko Kamakura, Masanori Matsuo, Harumi Ishii, Fumiko Mitsuboshi, and Yoriko Miura. Patterns of static prehension in normal hands. *The American journal of occupational therapy : official publication of the American Occupational Therapy Association*, 1980. 2

[58] John Russell Napier. The prehensile movements of the human hand. *The Journal of bone and joint surgery. British volume*, 1956. 2

[59] Steffen Puhlmann, Fabian Heinemann, Oliver Brock, and Marianne Maertens. A compact representation of human single-object grasping. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016. 2

[60] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20577–20586, 2022. 2

[61] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018. 2

[62] Vidit Jain and Erik Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. pages 577 – 584, 07 2011. 2

[63] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018. 2

[64] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 2

[65] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM Trans. Graph.*, 38(4), jul 2019. 2

[66] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 2

[67] Ravi Teja Mullapudi, Steven Chen, Keyi Zhang, Deva Ramanan, and Kayvon Fatahalian. Online model distillation for efficient video inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3573–3582, 2019. 2

[68] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[69] Ruoshi Liu, Sachit Menon, Chengzhi Mao, Dennis Park, Simon Stent, and Carl Vondrick. Shadows shed light on 3d objects. *arXiv preprint arXiv:2206.08990*, 2022. 3

[70] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017. 3

[71] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. 5

[72] Ruoshi Liu, Chengzhi Mao, Purva Tendulkar, Hao Wang, and Carl Vondrick. Landscape learning for neural network inversion. *arXiv e-prints*, pages arXiv–2206, 2022. 5, 16

[73] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 6, 12

[74] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 12

[75] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4332–4341, 2019. 12

[76] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 12

Appendix

A. Implementation details

A.1. Pre-trained Generative Models

Full-body pose prior. For our full-body generative model $\mathcal{P}$, we use the GitHub implementation of VPoser [33]. VPoser is a variational autoencoder [73] trained on data obtained by applying MoSh [13] on three publicly available human motion capture datasets: CMU [6], training set of Human3.6M [74], and the PosePrior dataset [15].

Hand object grasping model. For the right-hand grasping model $\mathcal{G}$, we use the GitHub implementation of GrabNet [19]. GrabNet consists of two networks: **1) CoarseNet** for coarse grasps and **2) RefineNet** for refining the coarse grasps. CoarseNet takes as input a latent vector w, and the object $\mathcal{O}$, represented by its BPS representation [75], and generates a hand pose, including its translation and rotation, to be used as input to MANO. RefineNet takes as input the CoarseNet grasp and the distances D from the coarse MANO vertices to the object mesh. It then refines the grasp through 3 iterations as in [76] to give the final grasp. RefineNet has been trained by sampling CoarseNet grasps as ground truth and perturbing the hand pose parameters to simulate noisy input estimates.

Object representation. The object to-be-grasped $\mathcal{O}$ is represented in GrabNet using the Basis Point Set (BPS) [75] representation which is capable of encoding arbitrary 3D object shapes. Given any object vertices, an approaching angle α and N_b fixed basis points, the BPS representation involves rotating the object by the angle α, placing it in the center of the points and calculating the minimum distance from each point to the nearest surface of the object. The outputs of our model are rotated by the inverse of the rotation matrix given by α. We use the implementation from the bps_torch library on GitHub.

A.2. Training Details

- For every example (scene and object), we optimize $N = 500$ latent vectors z, with different initializations, and at the end of the optimization process we select the ones that result in the smallest loss. During training, we periodically discard the 50% of the latent vectors that produce the largest losses. At the end of the optimization process, we end up with the best 16 samples out of the 500. The parameters of the mapping network (consisting of a 2-layer MLP) are shared across the N latent vectors.
- We constrain the value of w by normalizing it such that its norm is always one, following the density of a high-dimensional Gaussian prior, thus making sure w is within the distribution of the latent space of $\mathcal{G}$.
- We train with Adam optimizer with a learning rate of $1e-3$ for z, and $1e-4$ for the mapping network. Additionally, we found that the translation parameters have a much stronger gradient than the rest, especially the latent v, so we divide the gradient that goes through t_b^{xy} by 3, and multiply the gradient that flows through v by 10. We train for 500 iterations.
- Empirically, we found that the pre-trained GrabNet model was much more sensitive to approaching angle α then it was to the latent w, hence we set w to the zero vector in our experiments.
- For the hand matching loss, we weigh the vertices around the wrist more ($\times 3$) than the rest, as the alignment around the wrist is less noisy than in the fingers.
- The values for the loss weights λ in the total loss are set to: $\lambda_{\text{hm}} = 20$, $\lambda_{\text{o}} = 1000$, $\lambda_{\text{g}} = 0.01$. These do not necessarily reflect the importance given to each loss, as the loss values are in completely different scales.
- We additionally found that scaling the output of the MLP differently for every parameter was helpful. Specifically, we scale v by 5 (giving more flexibility to the human pose generator), the translation parameters by 10, and the angle and orientation by 20.
- We found that some obstacles have very thin walls, so we make the obstacle mesh $\mathcal{M}_{\text{obstacle}}$ thicker by 5mm, which allows us to model the intersections better.

B. ReplicaGrasp Dataset

Receptacles. We use a total of 48 receptacles from the ReplicaCAD dataset [34]. Some of the static rigid object receptacles include: apartment chair, sofa, table top, TV stand and wall cabinet. The receptacles from articulated objects include: refrigerator top, middle and bottom; top, middle and bottom drawers of kitchen counter on both right and left sides, and kitchen sink; as well as top, middle and bottom compartments of both sides of the kitchen cupboard. Many of the receptacles are visible in Fig. 1.

Objects. We obtain 50 everyday object meshes from the GRAB dataset and use the Habitat Simulator [35] to get the final locations of the objects on the receptacles. We use the GitHub v0.2.2 release. The simulator runs dynamics for 5 seconds to check for stability of newly placed objects.

C. Quantitative Analysis

We conduct a detailed analysis of the results in Table 1.

Performance as a function of object height. We demonstrate the need of having a benchmark like ReplicaGrasp

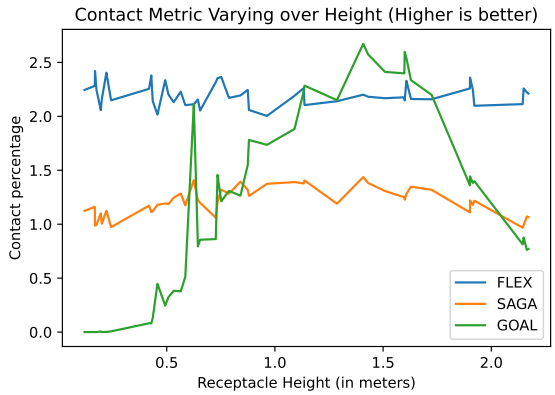


Figure 11. **Object contact percentage varying by height.** FLEX performs more consistently than both baselines and is best on average.

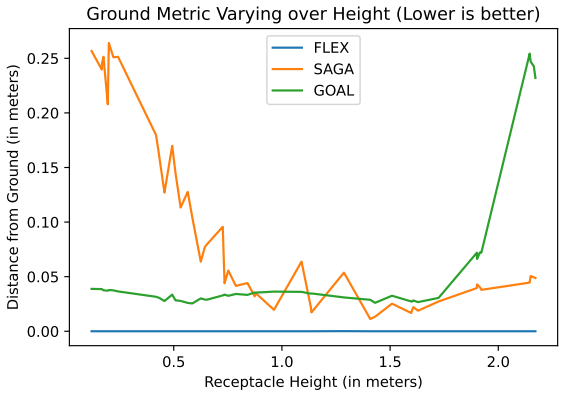


Figure 12. **Ground distance varying by height.** SAGA performs worst at lower heights while GOAL performs worst when the objects are high up.

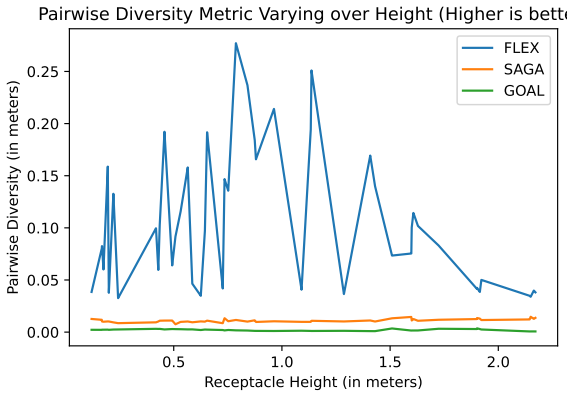


Figure 13. **Diversity varying by height.** GOAL and SAGA both consistently fail at generating diverse outputs at all heights. FLEX can generate most diverse grasps at medium heights.

that allows evaluating grasps at different heights. In GRAB, the object heights varies from a minimum of 0.75 meters to a maximum of 1.38 meters, with a mean of 1 meter. In ReplicaGrasp, our object heights have a much larger range from a minimum of 0.12 meters to a maximum of 2.2 meters, with a mean of 1 meter. We show how performance along different metrics changes by varying the heights of the objects.

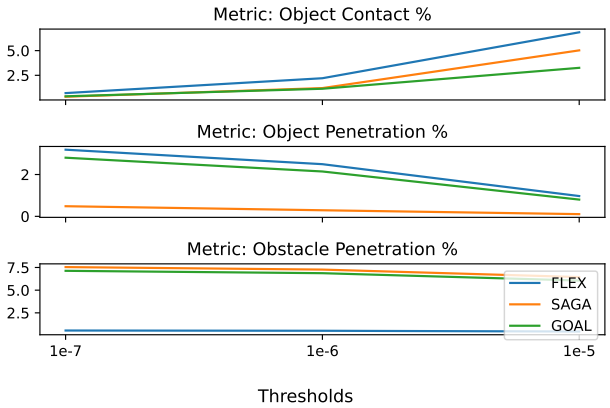


Figure 14. **Sensitivity to threshold σ.** Trends stay the same for all metrics across different thresholds.

- **Object contact percentage** - Fig. 11 shows that GOAL performs well when objects are at heights that have been seen during training, but sees drops in performance at other heights. SAGA is more consistent than GOAL even at varying heights. FLEX outperforms both baselines on average across all heights without showing much variation across height changes.
- **Ground distance** - Fig. 12 shows that when the object is at a low height, SAGA fails by generating humans with legs buried below the ground. SAGA is better at higher heights, although qualitatively the humans appear elongated. GOAL generates humans that try to fly up to grasp objects at larger heights. GOAL performs better at lower heights, although qualitatively the humans look unnatural with awkwardly bent legs.
- **Sample diversity** - In Fig. 13, we show the average pairwise diversity across pairs of samples generated for an instance, averaged across all instances of the dataset. This quantifies the method's ability to generate a range of complex human poses. FLEX outperforms both baselines by a large margin despite having additional constraints of avoiding obstacles.

Sensitivity of the metrics to the threshold. In order to compute the metrics, we set a threshold σ that determines the boundary between contact and penetration with an object or an obstacle. In Fig. 14 we report results of our metrics for different values of this threshold, and show that the metrics are not sensitive to its value, and that the trends shown in the main paper (where we use $\sigma = 1e-6$) hold.

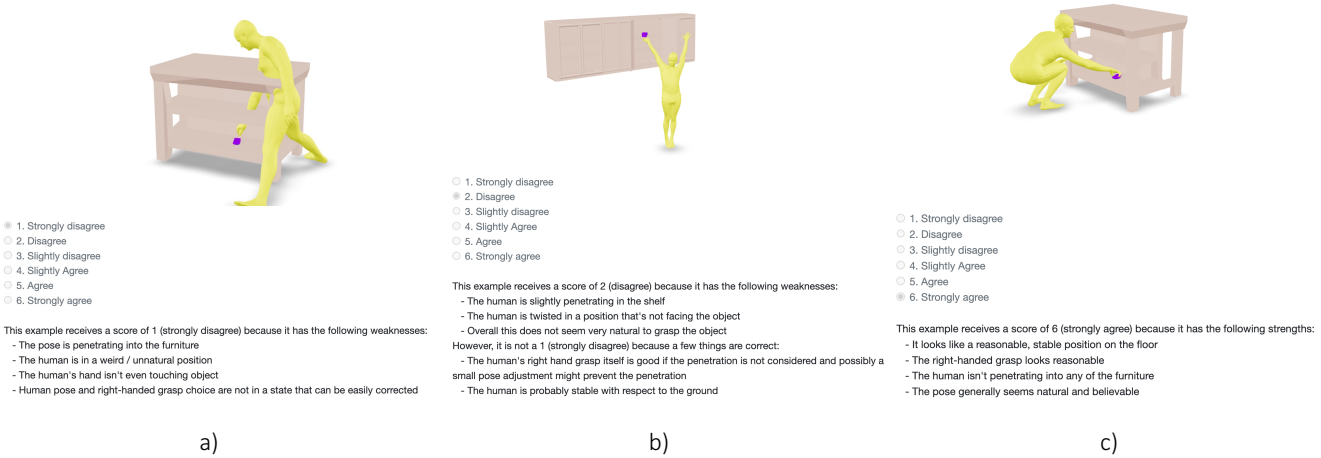


Figure 15. **Examples shown to Amazon Mechanical Turk subjects**. We provide these three examples to the subjects, which range from very bad (strongly disagree with the statement) to very good (strongly agree) results.

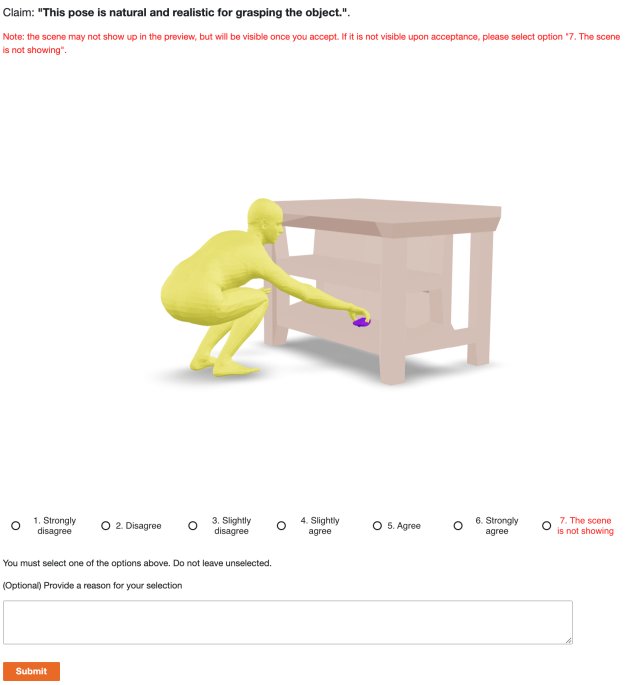


Figure 16. **Human studies.** Example of a HIT shown to subjects on Amazon Mechanical Turk.

D. Evaluation Metrics

We provide equations for each of the metrics described in Section 5.3.

Object contact percentage.

$$M_{\text{obj}}^{\text{contact}} = \frac{100}{|\mathcal{V}_{\text{o}}|} \sum_{i=1}^{|\mathcal{V}_{\text{o}}|} \mathbb{1}\left(|d_{\text{vm}}(\mathcal{V}_{\text{o}_i}, \mathcal{M}_{\text{human}})| \leq \sigma\right), \quad (5)$$

where $\mathcal{V}_{\text{o}}$ are the vertices of the object, $\mathcal{M}_{\text{human}}$ is the human mesh, d_{vm} is the signed vertex-to-mesh distance, σ is a small threshold and $\mathbb{1}$ is the indicator function.

Object penetration percentage.

$$M_{\text{obj}}^{\text{penet}} = \frac{100}{|\mathcal{V}_{\text{o}}|} \sum_{i=1}^{|\mathcal{V}_{\text{o}}|} \mathbb{1}\left(d_{\text{vm}}(\mathcal{V}_{\text{o}_i}, \mathcal{M}_{\text{human}}) < -\sigma\right) \quad (6)$$

Obstacle penetration percentage.

$$M_{\text{obst}}^{\text{penet}} = \frac{100}{|\mathcal{V}_{\text{b}}|} \sum_{i=1}^{|\mathcal{V}_{\text{b}}|} \mathbb{1}\left(d_{\text{vm}}(\mathcal{V}_{\text{b}_i}, \mathcal{M}_{\text{obstacle}}) < -\sigma\right), \quad (7)$$

where $\mathcal{V}_{\text{b}}$ are the vertices of the human body and $\mathcal{M}_{\text{obstacle}}$ is the obstacle mesh. In contrast to Eq. (6), here we average over the human body (not the obstacle) vertices, because we care about how much of the human is penetrating an obstacle, not how much of the obstacle is being penetrated by the human.

Ground distance.

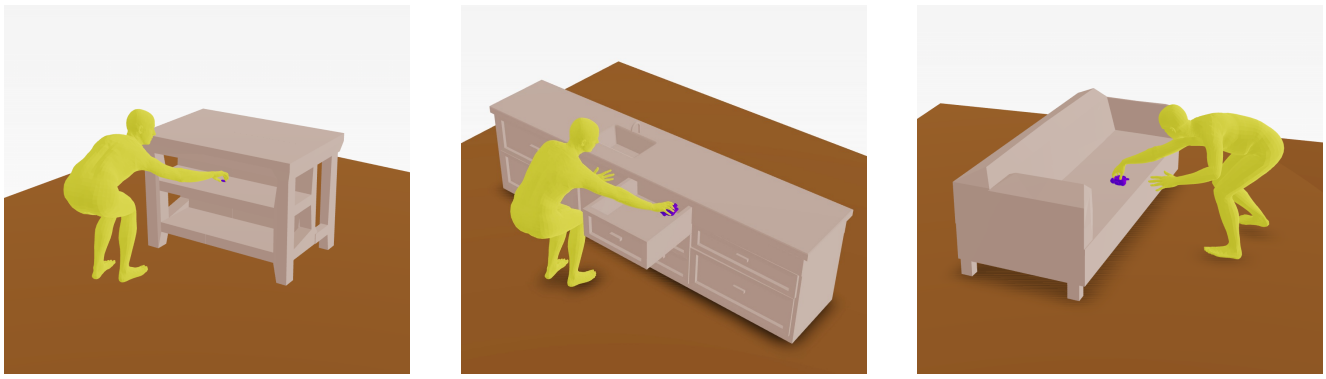
$$M_{\text{ground}} = \left|\min(\mathcal{V}_{\text{b}}^{z})\right|, \quad (8)$$

where $\mathcal{V}_{\text{b}}^{z}$ is the z component of all vertices in $\mathcal{V}_{\text{b}}$.

Sample diversity.

$$M_{\text{Div}_{\text{samp}}} = \frac{2}{\mathcal{N}_{\text{s}} \cdot (\mathcal{N}_{\text{s}} - 1)} \sum_{\substack{i,j \in \mathcal{N}_{\text{s}} \\ i \neq j}} d_{\text{vv}}(\mathcal{V}_b^{(i)}, \mathcal{V}_b^{(j)}), \quad (9)$$

where $\mathcal{N}_{\text{s}}$ is the number of samples for a single example and d_{vv} is the L^2 vertex-to-vertex distance in the 3D space. $\mathcal{V}_b^{(i)}$ represents the human body vertices corresponding to the i-th sample.



a) Strongly agree b) Agree c) Slightly disagree

Figure 17. **Examples of human ratings.** The figure shows three samples and their corresponding human ratings. See Appendix E for comments from subjects.

Overall diversity.

$$M_{\text{Div}_{\text{all}}} = \frac{2}{\mathcal{N}_{\text{d}} \cdot (\mathcal{N}_{\text{d}} - 1)} \sum_{\substack{i,j \in \mathcal{N}_{\text{d}} \\ i \neq j}} d_{\text{vv}}(\mathcal{V}_b^{(i)}, \mathcal{V}_b^{(j)}), \qquad (10)$$

where $\mathcal{N}_{\text{d}}$ is the total number of instances in the dataset.

E. Human Studies

We conducted perceptual evaluation studies on Amazon Mechanical Turk (AMT) with a prompt as shown in Fig. 16. The specific instructions were as follows:

```
We want to evaluate the realism of the
    humans. Some questions to ask yourself
    while solving the task:
1) Would you expect to see a human like
    this in real-life?
2) Is the hand grasp going to result in a
    natural grasp?
3) Is the human stable on the ground?

The scene can be navigated by:
1) Clicking and dragging the mouse, to
    rotate the scene.
2) Zooming in and out with the scroll
    wheel.
3) Clicking at a point in the scene, to
    position that point in the center of
    the scene.

See examples for a better intuition.
```

Further, we showed subjects three examples of how to successfully perform the task by showing an example that deserves the ratings of 1, 2 and 6 respectively with explanations for the reasoning as shown in Figure 15. We randomly selected 96 examples of ReplicaGrasp covering objects in all 48 receptacles in both upright and fallen orientations. We showed each example to 5 different subjects and we had 30 unique subjects solve the task. We filtered out cases which saw high inter-subject disagreement.

	Sample-wise ↑		Overall ↑	
Method	Full Body	Right-hand	Full Body	Right-hand
GOAL	0.11	0.04	6.01	12.14
SAGA	1.14	0.09	15.29	13.79
FLEX (ours)	**26.91**	**0.36**	**39.98**	**16.40**

Table 3. Diversity analysis on GRAB (cm)

Fig. 17 shows some examples of FLEX generations evaluated by subjects. Participants generally found the results realistic – for example, for Fig. 17 b, a participant wrote: *"The stretch of the hand inside the drawer is very realistic"*. In some cases where the subjects gave a low rating, for instance in Fig. 17 c, we received interesting comments: *"Doesn't need to squat to grab item"*. This reveals a shortcoming of our system wherein we do not measure the effort required to grasp an object. Explicitly modeling physical effort and its effect on the choice of the human's pose is an interesting direction that we leave as future work.

F. Computational Budget

We performed speed and memory comparisons (averaged across 10 runs) for generating 16 different samples on a single RTX 2080 Ti GPU. FLEX involves using pretrained models simultaneously, the memory consumption is 3x (4.8 GB vs 1.4 GB). FLEX takes around 8.5 minutes to generate 16 samples, while SAGA and GOAL take 6 and 1 minute respectively. We sacrifice computational budget for significantly better results.

G. Diversity Analysis

Tab. 3 shows diversity metric computed for hand (no-full-body) and for full-body (no-hand) for all 3 methods. FLEX has higher diversity in both, but the gains are significant for full-body.

Method	Obj Cont (%) ↑	Obj Penet (%) ↓	Obs Penet (%) ↓	Ground (cm) ↓
Random Init	0.15	35.06	2.02	60.32
CMA-ES	0.03	17.39	2.03	58.25
FLEX	**2.20**	**2.50**	**0.53**	**0.00**

Table 4. Comparison with Optimization methods.

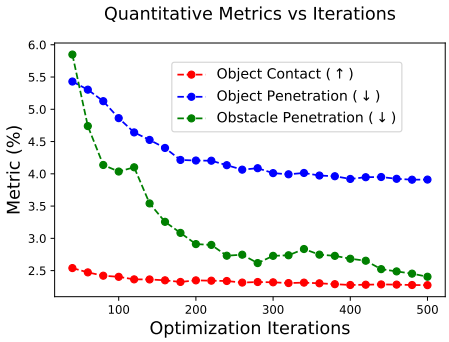


Figure 18. Object contact, penetration and Obstacle penetration metrics varying by iterations.

H. Choice of Optimization Framework

FLEX is agnostic to the choice of the optimization method. We used the recent Liu *et al.* [72] which smooths the loss landscape for better convergence. To validate this choice of a gradient-based optimization framework, we conduct experiments with non-gradient based methods described below:

- **Ranking:** Instead of optimization, we simply rank a large the batch of whole-body grasps produced by randomly sampling the optimization parameters.
- **CMA-ES:** Covariance matrix adaptation evolution strategy implemented from PyPI.

Results are shown in Tab. 4. As expected, FLEX is superior to both the baselines.

I. Performance as a function of the number of optimization steps

Fig. 18 shows average optimization metrics over different iterations. Object and obstacle penetration go down with training. Object contact stays largely unchanged.