G-HOP: Generative Hand-Object Prior for Interaction Reconstruction and Grasp Synthesis

Yufei Ye[1] Abhinav Gupta[1] Kris Kitani[1,2] Shubham Tulsiani[1]
[1]Carnegie Mellon University [2]MetaAI
https://judyye.github.io/ghop-www

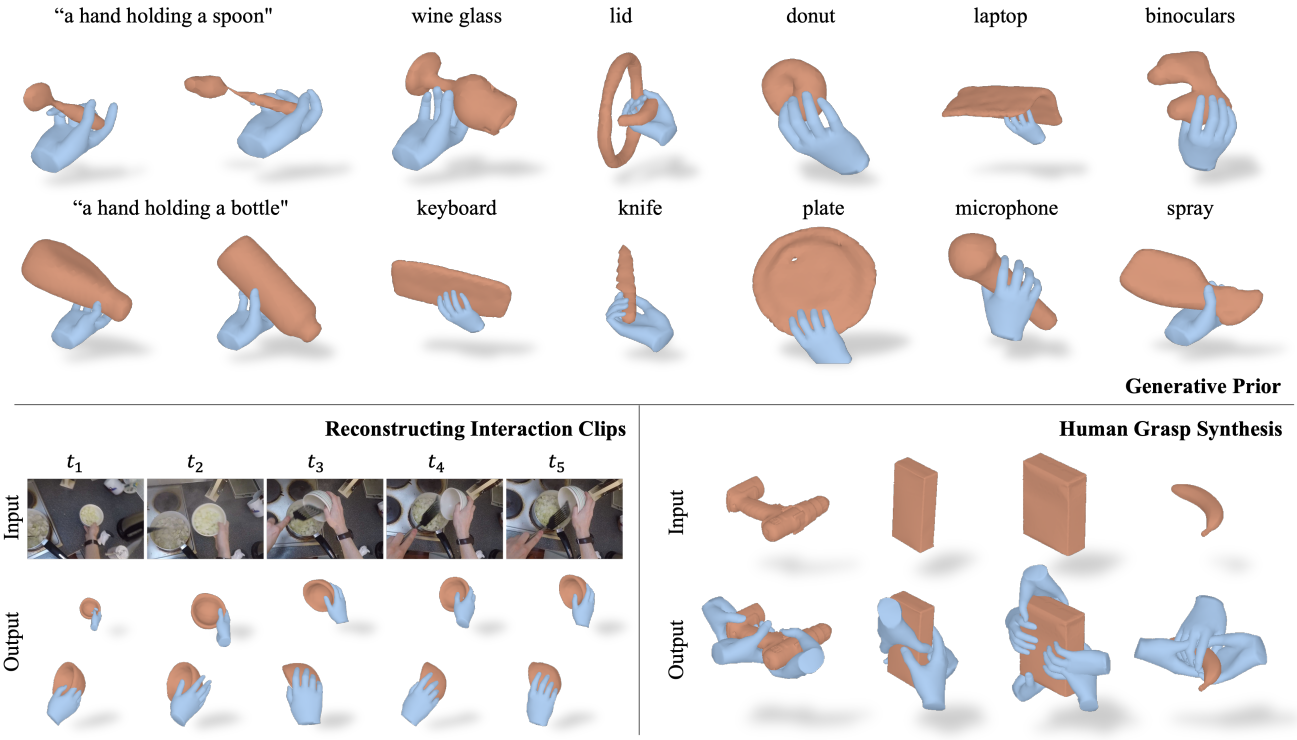


Figure 1. G-HOP can generate plausible hand-object interactions across a wide variety of objects (top). The learned generative prior can also guide inference for tasks such as reconstructing everyday interaction clips and synthesizing human grasps given object meshes.

Abstract

We propose G-HOP, a denoising diffusion based generative prior for hand-object interactions that allows modeling both the 3D object and a human hand, conditioned on the object category. To learn a 3D spatial diffusion model that can capture this joint distribution, we represent the human hand via a skeletal distance field to obtain a representation aligned with the (latent) signed distance field for the object. We show that this hand-object prior can then serve as generic guidance to facilitate other tasks like reconstruction from interaction clip and human grasp synthesis. We believe that our model, trained by aggregating seven diverse real-world interaction datasets spanning across 155 categories, represents a first approach that allows jointly generating both hand and object. Our empirical evaluations demonstrate the benefit of this joint prior in video-based reconstruction and human grasp synthesis, outperforming current task-specific baselines.

1. Introduction

Imagine holding a bottle, or a knife, or a pair of scissors. Not only can you picture the differing shapes of these objects *e.g.* a cylindrical bottle or a flat knife, but you can also easily envision the varying configurations your hand would adopt when interacting with each of them. Even though the form of these hand-object interactions may vary widely depending on factors such as geometry (*e.g.* we will hold a pen

and a pan rather differently), or intent (*e.g*. passing a knife vs. using it to cut), we humans can effortlessly picture such interactions with everyday objects in our daily lives. In this work, our goal is to build a computational system that can similarly generate plausible hand-object configurations.

Specifically, we learn a denoising diffusion-based generative model that captures the joint distribution of both hand and object during interaction in 3D. Given a category-conditioned description *e.g*. 'a hand holding a plate', our generative model can synthesize both, plausible object shape as well as the relative configuration and articulation of the human hand (see Fig. 1 top). A key question we address is that what are good HOI *representations* for the model. While objects shapes are typically described via spatial (signed) distance fields, human hands are commonly modeled via a parametric mesh controlled by an articulation variable. Instead of modeling these disparate representations in our generative model, we propose a homogeneous HOI representation and show that this allows learning a 3D diffusion model that jointly generates the hand and object.

In addition to enabling synthesis of diverse plausible hand and object shapes, our diffusion model can also serve as a generic prior to aid inference across tasks where such a representation is a desired output. For example, the ability to reconstruct or predict interactions is of central importance for robots aiming to learn from humans, or virtual assistant trying to aid them. We consider two well-studied tasks along these lines: i) reconstructing 3D hand-object shapes from everyday interaction clips, and ii) synthesizing plausible human grasps given an arbitrary object mesh. To leverage the learned generative model as a prior for inference, we note that our diffusion model allows computing the (approximate) log-likelihood gradient given any hand-object configuration. We incorporate this in an optimization framework that combines the prior likelihood-based guidance with task-specific objectives (*e.g*. video reprojection error for reconstruction) or constraints (*e.g*. known object mesh for synthesis) for inference.

While understanding hand-object interactions is an increasingly popular research area, real-world datasets capturing such interactions in 3D are still sparse. We therefore aggregate 7 diverse real-world interaction datasets resulting in long-tailed collection of interactions across 157 object categories, and train a shared model across these. To the best of our knowledge, our work represents the first such generative model that can jointly generate both, the hand and object, and we show that it allows synthesizing diverse hand-object interactions across categories. Moreover, we also empirically evaluate the prior-guided inference for the tasks of video-based reconstruction and human grasp synthesis, and find that our learned prior can help accomplish both these tasks, and even improve over task-specific state-of-the-art methods.

2. Related Works

Reconstructing Hand-Object Interactions. Reconstructing HOI interactions from images or videos can be challenging due to heavy mutual occlusions, and several initial approaches [3, 14, 16, 45] simplified the task by requiring an instance-specific object template and reducing the task to 6D pose estimation. Some recent video-based reconstruction methods [17, 21, 49] show promising results without requiring templates, but they target in-hand scanning setups where abundant multi-view cues are available and cannot infer unobserved regions. Another line of template-free methods [6, 7, 18, 24, 37, 51] uses data-driven prior for reconstructing general objects from single images, but these are not temporally-consistent given input videos. Most closely related to our work is DiffHOI [52] which leverages both multi-view cues and data-driven priors via per-sequence optimization. We adopt this framework and show the our proposed generative 3D prior can yield better reconstruction, while also enabling inference across other tasks.

Grasp Synthesis. Grasp synthesis studies how to interact with an objects plausibly. A line of work pursues 2D representations of interactions, or visual affordance. Given a 2D image, they predict interactions in various forms like trajectory, heatmaps, keypoints, or synthesized images [13, 30, 35, 53]. However, interaction represented in 2D can not be directly used to command a robot to grasp an object in 3D. There are extensive works in robotics that predict 3D robot grasp [1, 2, 27, 33] for different end-effectors. Meanwhile, human grasp as a special end-effector receives great attention [3, 12, 15, 22, 24, 26, 31]. Most relevant work including GF [24] and GraspTTA [22] model a conditional probability of human hand given an object mesh. In contrast to the task-specific methods, we directly leverage the generic joint hand-object generative prior and show that this leads to more natural human grasps.

Diffusion Models as Generative Prior. Diffusion models [20] are a family of generative models and have driven great progress in multiple domains like image generation [39, 40], 3D object generation [23, 28, 36], novel-view synthesis [29, 34], human motion [25, 46], video generation [43], *etc*. An advantage of diffusion models is that they allow computing log-likelihood gradients via score distillation [36, 48] and thus can be used as foundation generative priors for other tasks [11, 34, 42, 54]. In our work, we use diffusion model to learn a generative prior for 3D hand-object interactions and apply it to the tasks of HOI reconstruction and grasp synthesis.

3. Method

We first seek to model the joint distribution of the geometry of hand-object interactions $p(\mathbf{O}, \mathbf{H}|\mathbf{C})$ where $\mathbf{C}$ is the text

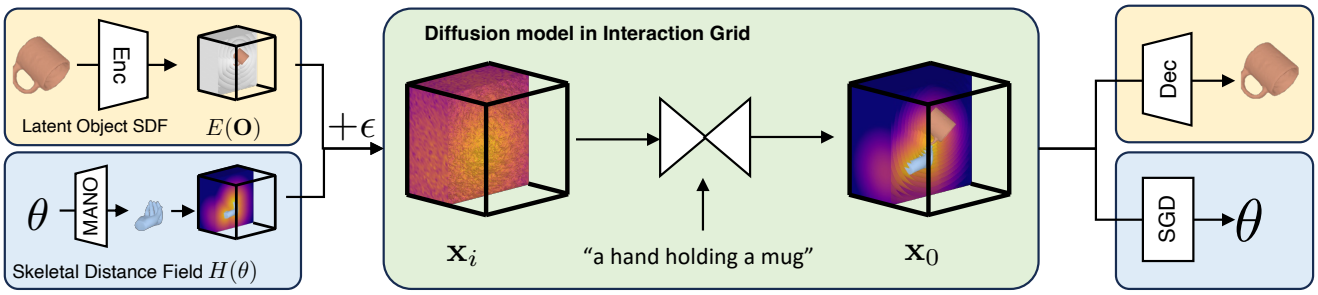


Figure 2. **Method Overview of Generative Hand-Object Prior:** Hand-object interactions are represented as interaction grids within the diffusion model. This interaction grid concatenates the (latent) signed distance field for object and skeletal distance field for the hand. Given a noisy interaction grid and a text prompt, our diffusion model predicts a denoised grid. To extract 3D shape of HOI from the interaction grid, we use decoder to decode object latent code and run gradient descent on hand field to extract hand pose parameters.

of an object category. We use a diffusion model Ψ to learn this generative prior, and propose a spatial interaction grid representation for learning (Sec. 3.1). We then apply this learned prior to guide reconstruction from monocular video clips and human grasp synthesis (Sec. 3.2). For both tasks, we frame inference as test-time optimization that combines task-specific constraints/objectives with score "distillation" from the pre-trained diffusion model.

3.1. Generative Hand-Object Prior

In this work, we propose 'interaction grids' as a homogeneous HOI representation that allows the diffusion models to effectively reason about the 3D hand-object interactions. Specifically, an interaction grid (Fig. 2) is a concatenation of a latent signed distance value grid representing the object $E(\mathbf{O})$ and a 'skeletal distance' field based grid parameterized by 3D hand pose $H(\boldsymbol{\theta})$, *i.e.* $\mathbf{x} \equiv (E(\mathbf{O}), H(\boldsymbol{\theta}))$. We model the interaction grid in a normalized hand-centric frame, where the hand palm always faces upwards. The hand-centric frame more effectively captures the inherent structures of interaction common to various objects, such as grasping handles, regardless of whether the object is a kettle or a power drill [51].

Latent Object Signed Distance Field. We use a signed distance field (SDF) grid to capture object details. As the memory grows cubically with grid resolution, we follow prior works to use a VQ-VAE [47] to compress high-resolution SDF grids into lower-dimension object latent. $\mathbf{z} = E(\mathbf{O}), \mathbf{O} = D(\mathbf{z})$. Note that when training the autoencoder, the object SDF grids are also transformed into hand-centric frame.

Skeletal distance field for Parametric Hand. While there is consensus on how to represent objects, it is unclear what is a good representation of hand during interaction. Many prior works generate hand/human shape by diffusing in the compact pose parameter space [25, 46] but we find this space not ideal when we diffuse it jointly with objects latent grids (see supplementary) probably because the diffusion model cannot easily to reason about spatial interactions using this heterogeneous representation (1D articulation vector and 3D SDF grid). Instead, we propose to represent hand in a pose-parameterized distance field $H(\boldsymbol{\theta})$. It is a 15-channel 3D grid that encodes the distance to each joint. $H(\boldsymbol{\theta})[u, v, w]_{i=1:15} \equiv \|\mathbf{X}_{[u,v,w]} - J_i\|_2^2$. This skeletal distance field can be converted from pose parameter space and vice versa by leveraging differentiable parametric mesh model MANO [41]. Specifically, MANO takes in the pose parameter and outputs joint position $J_i(\boldsymbol{\theta})$ to compute the skeletal field. To recover pose parameter $\boldsymbol{\theta}$ from a skeletal distance field, we run gradient decent on pose parameter to minimize the distance between the induced field and the given field, $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}}(H(\boldsymbol{\theta}) - \hat{H}) + w\|\boldsymbol{\theta}\|_2^2$.

Denoising Diffusion Model. In training, the diffusion model takes in a text embedding and a noisy 3D interaction grid $\mathbf{x}_i$ and is supervised to restore the clean grid $\hat{\mathbf{x}}_0$.

$$\mathcal{L}_{\text{DDPM}}[\mathbf{x}; \mathbf{C}] = \mathbb{E}_{i,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} w_i \|\hat{\mathbf{x}}_0 - \Psi(\mathbf{x}_i, i, \mathbf{C})\|_2^2 \quad (1)$$

The object distance field is in resolution 64^3 and the VQ-VAE downsamples the resolution to 16^3 which is then concatenated with the hand skeletal field. We implement the diffusion model as 3D-UNet with three 3D convolution blocks. The text prompt is encoded by CLIP [38] text encoder and is passed to the 3D-UNet by cross-attention at each block.

3.2. Prior-guided Reconstruction and Generation

Given the learned generative prior, we leverage it for both HOI reconstruction and human grasp synthesis. The inference in both tasks is performed via test-time optimization which is guided by distilling the learned prior. We use score distillation sampling (SDS [36, 48]) to approximate log-probability gradients of interaction grids $\mathbf{x}$ from the diffusion model. Specifically, to guide the grid to be more plausible at every optimization step, we corrupt the current interaction grid $\mathbf{x}$ by a certain amount of noise and let diffusion model denoise it. The discrepancy between this

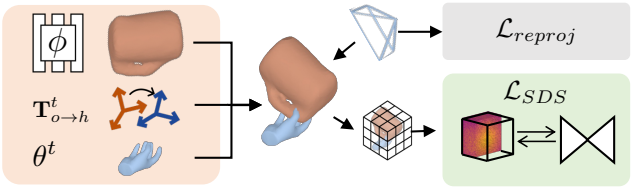


Figure 3. **Reconstructing Interaction Clips:** We parameterize HOI scene as object implicit field, hand pose, and their relative transformation (left). The scene parameters are optimized with respect to the SDS loss on extracted interaction grid and reprojection loss (right).

denoised prediction and the current estimate can be be used an objective to obtain log-likelihood gradients:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}) \approx \nabla_{\mathbf{x}} L_{SDS}[\mathbf{x}] = \mathbb{E}_{\epsilon,i}[w_i(\mathbf{x} - \hat{\mathbf{x}}_i)] \quad (2)$$

In the following section, we will show that both reconstruction and grasp synthesis can leverage the common optimization frameworks by instantiating task-specific parameters and constraints.

3.2.1 Reconstructing Interaction Clips

Given a video clip depicting a hand interacting with a rigid object, we aim to reconstruct the underlying 3D shape of the hand and the object. We follow DiffHOI [52] which performs inference via a optimizing 3D scene representation with respect to a reprojection term and a data-driven prior term. Instead of their 2D diffusion prior which can only guide object shape inference, we substitute our learned joint 3D generative prior and show that it leads to improved performance for video-based reconstruction.

Scene Parameters and Rendering. We adopt a similar representation as DiffHOI [52], which decomposes the HOI scene into three parts: i) a time-persistent object signed distance field represented by an implicit neural network $\phi(\cdot)$; ii) time-varying hand pose parameters $\boldsymbol{\theta}^t$, and iii) the relative poses $\mathbf{T}^t_{o\to h}$ between them. This scene representation can be rendered into 2D masks $\mathbf{I}^t$ by differentiably compositing renderings of the volumetric object and hand mesh.

Prior-Guided Reconstruction. Different from DiffHOI, our data-driven prior is in 3D space instead of 2D. Furthermore, our prior also models the hand pose rather than use it as a condition, and can thus also provide gradients to guide hand pose optimization. Specifically, to regularize the 3D representation, we query the 3D volume in the hand-centric frame to get interaction grid for each frame and pass the grid to the pre-trained diffusion model, *i.e.* $\mathbf{x}^t = (E(\phi(\mathbf{T}^{t^{-1}}_{o\to h} X_{grid})), H(\boldsymbol{\theta}^t))$, where X_{grid} is the coordinate of the queried volume. Other losses are similar to [52]: the reprojection term is computed in the mask space

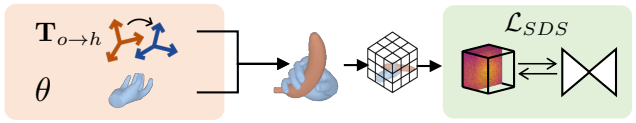


Figure 4. **Grasp Synthesis:** We parameterize human grasps via hand articulation parameters and the relative hand-object transformation (left). These are optimized with respect to SDS loss by converting grasp (and known shape) to interaction grid (right).

$\mathcal{L}_{reproj} = \|\mathbf{I}^t - \hat{\mathbf{I}}^t\|$; other regularization include Eikonal loss and temporal smoothness.

The optimization converges faster than previous work, perhaps because the prior in 3D provides stronger supervision. Specifically, we optimize 15000 iterations for each video clips which takes about an hour (which is 85% faster than DiffHOI [52]).

3.2.2 Synthesizing Plausible Human Grasps

Given an object mesh M_o, we aim to synthesize human grasps for the object. Formally, this corresponds to sampling from the conditional distribution $p(\mathbf{H}|\mathbf{O}, \mathbf{C})$. While our diffusion model captures the joint distribution of hand and object, it does not allow sampling human grasp directly given an object. Instead, we obtain plausible grasps via a test-time optimization approach to seek grasping modes while constraining the object to match the input. We also provide a mechanism to rank the generation by measuring consensus between diffusion model and the grasp synthesis.

Grasp Parameters. We parameterize a human grasp by the relative pose of the hand with respect to the object $\mathbf{T}_{o\to h}$, along with its articulation $\boldsymbol{\theta}$. We initialize hand articulation to a mean configuration while initializing relative pose with a random orientation and translation.

Optimization. In order to use diffusion model to guide grasp synthesis, we first convert the object mesh into SDF grid G_o, which is then transformed from the object-centric to the diffusion model coordinate (hand-centric) by the relative pose $\mathbf{T}_{o\to h}$, *i.e.* $\mathbf{x} = (E(\mathbf{T}_{o\to h} G_o), H(\boldsymbol{\theta}))$. We optimize the relative pose along with hand articulation for 500 iterations by maximizing the interaction likelihood from Eq. 2, *i.e.* $\log p(\boldsymbol{x}(\boldsymbol{T}_{o\to h}, \boldsymbol{\theta}))$. To account for accuracy loss when converted to low-resolution grids, we refine the predicted hand with the original mesh to encourage surface contact and penalize mesh collision. We show in supplementary that the distillation provides a good initialization for the mesh refinement while surface refinement further improves contact and grasp stability.

Ranking Grasps. The proposed approach to grasp synthesis is stochastic due to different initialization and the stochastic distillation process. Thus diverse grasps can be sampled. Furthermore, many applications like robotic manipulation would also want to know how plausible each

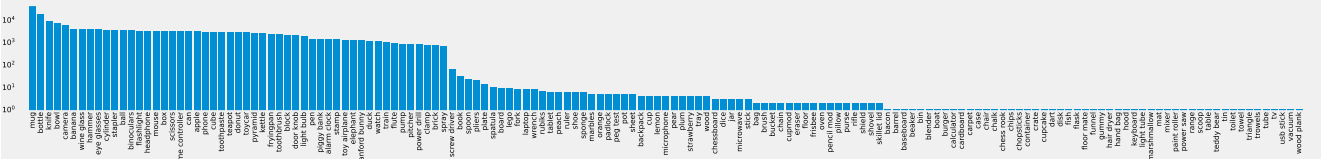


Figure 5. **Dataset Statistics:** number of training samples for each category when training our generative prior. Zoom in for better view.

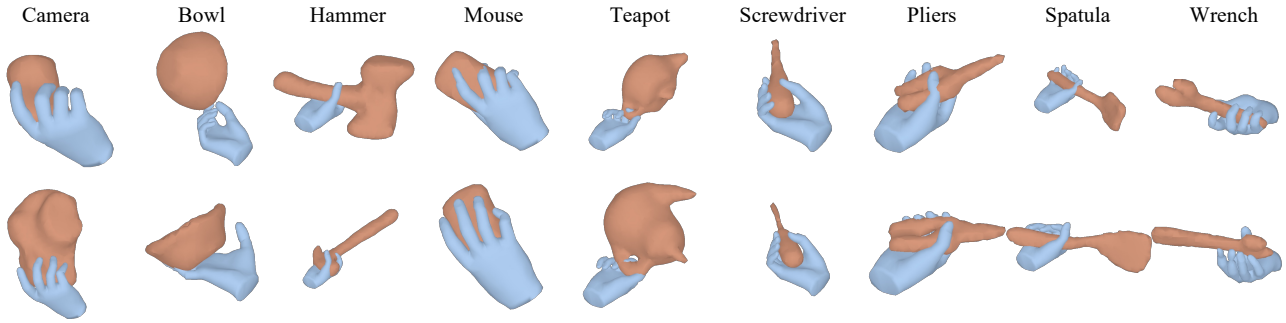


Figure 6. **Generations from Generative Hand-Object Prior:** Given a text prompt (only showing class label), we visualize two generated interactions from G-HOP . Categories are sorted from most common to least common in training (left to right). Generations are diverse in terms of object shape such as teapots, hand articulation such as mouse, and use intent like hammer.

grasp is. We also propose a mechanism to evaluate the sampled grasp. We approximate the likelihood upper bound [19] by averaging SDS loss across different time steps i:

$$s(\theta, \mathbf{T}_{o\to h}) = -\sum_{i=1}^{T} w_i \|\mathbf{x}(\boldsymbol{\theta}, \mathbf{T}_{o\to h})) - \hat{\mathbf{x}}_i(\epsilon)\|_2^2 \quad (3)$$

Intuitively, this measures the agreement between the prediction and the denoised output from the diffusion model, which indicates the distance of the current grasp to a plausible mode. We observe that this score provides a consistent and meaningful ranking across different samples.

4. Experiments

We train the generative prior on a collection of HOI datasets. We first show data distribution on this dataset collection and then visualize samples from the learned generative prior (Sec 4.1). In Sec. 4.2, we show that the learned prior benefits the task of reconstructing interaction clips. Our method outperforms other reconstruction baselines on HOI4D and we also show reconstruction of in-the-wild videos. In Sec. 4.3, we evaluate human grasps that are synthesized by directly applying our learned prior. We compare G-HOP to other baselines on two datasets and conduct user study to show that human grasp synthesized by ours is the most preferred one.

Training Data. We train our diffusion model on a combination of several world datasets including [3, 5, 8, 32, 44, 50], using their annotated 3D meshes of hand and objects. The name of categories across datasets are not standardized so we manually map synonyms or different formats to the same word (*e.g.* cellphone, iphone → phone, doorknob, door_knob → door knob). In total, we reduce 362 different words to 155 classes. All training data were converted into SDF grids, in hand-centric frame, with a resolution of 64^3 and spanning 30cm in all directions.

4.1. Visualizing Data-Driven Prior

We visualize the number of training samples per class in Fig. 5. The data is extremely unbalanced and follows a long-tail distribution. Classes with most training samples like mug consist of more than 10k grasps while few-shot classes such as skillet lid consist of fewer than 100 grasps.

In Fig. 6, we visualize hand-object interactions generated from the learned generative prior. We show 3 samples in different rows for each class. The classes from left to right are sorted by the training size from more to less. We see that the generated objects vary in shape. For example, different cameras display various lengths of lens. The generated samples are also diverse in terms of ways to hold them. Some hammers are held by handles and some are held by heads (for hand-over). We also find that the model can generate diverse and plausible samples on few-shot classes (shown on the right side).

4.2. Reconstructing Interaction clips

Setup and Evaluation Metrics. We evaluate interaction reconstruction on the HOI4D dataset. HOI4D is an egocentric dataset recording people interacting with different objects. We use the same split as DiffHOI [52] that consists of 2 video clips for all portable rigid object categories. The objects in the test set are held out from the

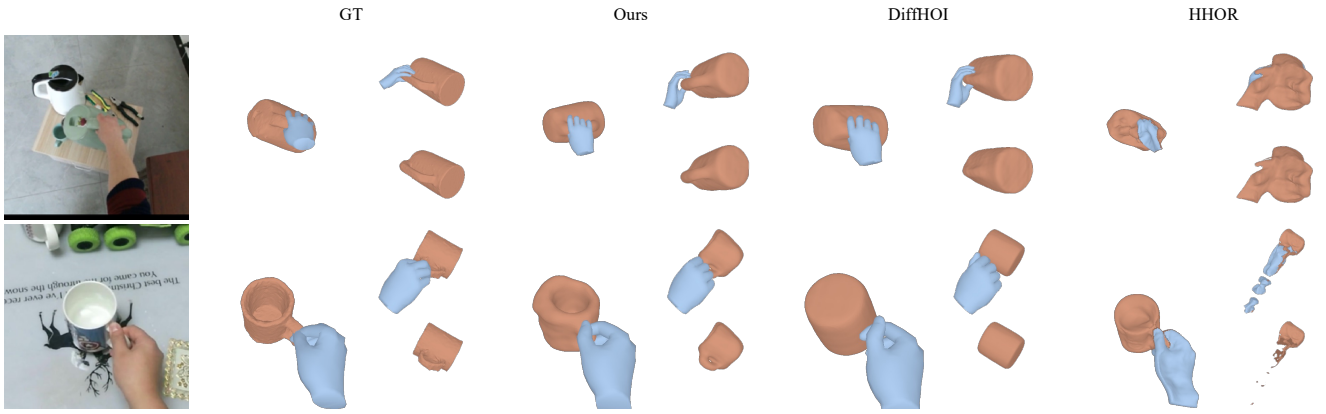


Figure 7. **Qualitative Evaluation on HOI4D:** We show reconstruction by G-HOP and two other video reconstruction baselines [21, 52] in the image frame (left) and from another view with (top right) or without (bottom right) reconstructed hand. Please see our project page for reconstruction videos from all methods.

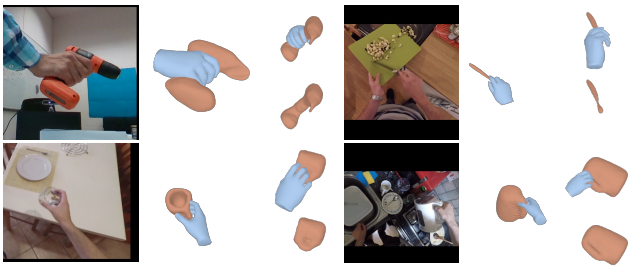


Figure 8. **In-the-Wild Reconstruction:** reconstruction on interaction clips from novel datasets [10, 16].

train set. We evaluate three aspects of the output: object reconstruction error, hand reconstruction error (MPJPE, AUC), and hand-object alignment (CD_h). Following prior works [21, 52], we align the object reconstruction with the ground truth by scaled Iterative Closest Points (ICP) and report F-score at 5mm, 10mm, and Chamfer distance in the aligned space. To evaluate the relation between hand and object, we report Chamfer distance of objects in hand-centric frame $CD_h \equiv CD(\mathbf{T}^t_{o\to h}O, \hat{\mathbf{T}}^t_{o\to h}\hat{O})$.

Baselines. We compare with three other template-free baselines that tackle reconstruction from casual monocular interaction clips.

i) *iHOI* [51] is a single-view 3D reconstruction method that learns to map from image feature and hand articulation to in-hand object shape. The model is finetuned on HOI4D and reconstruction is evaluated per-video frame.

ii) *HHOR* [21] optimizes a hand-object field with respect to the input video without any data-driven prior.

iii) *DiffHOI* [52] is closest to our work. The main difference is that the prior in their work takes hand pose as input thus modeling the *conditonal* probability $p(\pi(O)|\pi(H), C)$. Additionally, their prior is an image-based diffusion model instead of a 3D diffusion model.

iv) *G-HOP (Cond)* is our ablated models that is conditioned on hand pose and text prompt (same as DiffHOI but with 3D backbone). It aims to disentangle the effect of upgrading the prior from 2D to 3D from modeling joint instead of conditional probability.

For fair comparison, our diffusion model for HOI4D evaluation only trains on HOI4D train split. All other experiments use the model trained on all datasets.

Table 1. **Comparing HOI reconstruction:** object error (F@5mm, F@10mm, CD), hand-object alignment CD_h, and hand error (MPJPE, AUC) on HOI4D. We compare G-HOP with baselines and also ablate if reconstruction benefits from priors in the 3D space or from joint modeling hand and object.

	Object Error			Align	Hand Error	
	F5↑	F10↑	CD↓	CD_h ↓	MPJPE↓	AUC↑
iHOI [51]	0.42	0.70	2.7	27.1	1.19	0.76
HHOR [21]	0.31	0.55	4.7	165.4	-	-
DiffHOI [52]	0.62	0.91	0.8	48.7	1.12	0.78
G-HOP	**0.76**	**0.97**	**0.4**	**18.4**	**1.05**	**0.79**
G-HOP(Cond)	0.66	0.92	0.7	19.3	1.14	0.77

Results. We visualize reconstructions from different methods in Fig. 7 in the image frame and from a novel viewpoint. HHOR, which does not leverage data-driven learning, struggles with unobserved regions and outputs degenerate solutions as shown from the novel view. While iHOI reconstructs better shapes for each frame, there are not temporally consistent (shown in supplementary video) and it cannot benefit from multi-view cues. In comparison, DiffHOI reconstructs temporally consistent and more realistic results, but the reconstructed shape is relatively coarse. For instance, the kettle handle is merely a bump on top of a cylinder and the reconstruction does not reflect the concavity of the mug. In contrast, the reconstruction from G-HOP captures more details of object shape. In the bottom

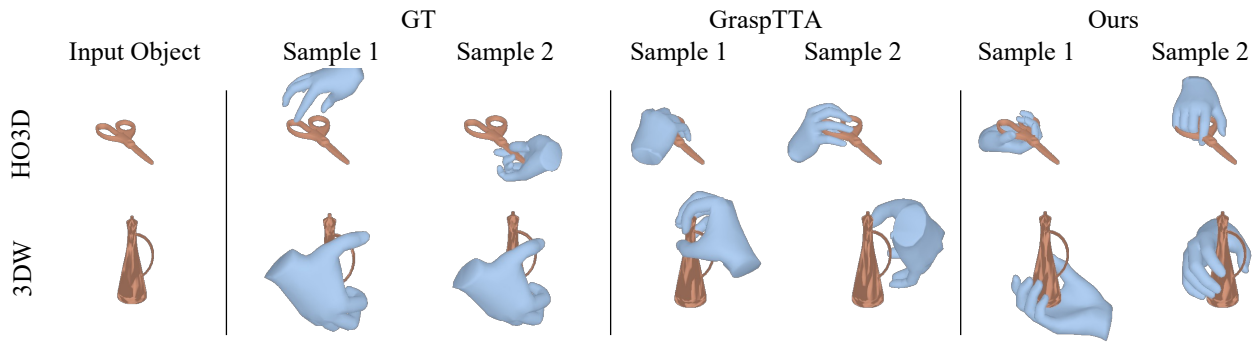


Figure 9. **Visualizing Grasp Generations:** Given an object mesh (left) from HO3D or 3DW, we sample two grasps from each method.

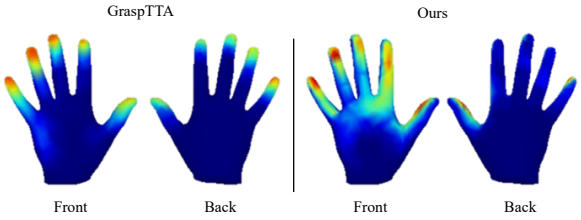


Figure 10. **Contact Map on Hand:** We visualize contact probability on hand over all generated samples from G-HOP and GraspTTA [22] on the HO3D dataset.

row, it even captures the space between the handle and the cup body. The visualization is consistent with the quantitative results in Tab. 1. Furthermore, we also find that the hand pose reconstruction also improves since the prior in G-HOP can also guide hand pose as well.

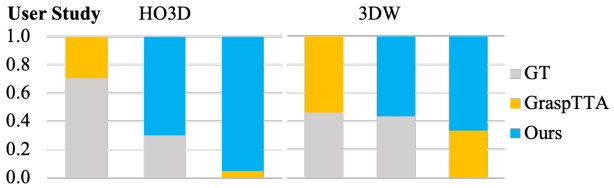
Ablations. Comparing with the ablated 3D conditional model (Tab. 1), we find that upgrading 2D prior to 3D improves object reconstruction significantly but does not improve hand reconstruction much. Joint modeling leads to better hand pose, which can in return improve object reconstruction further. Interestingly, we also find that the variant that *jointly* models HOI in *image* space performs even worse than DiffHOI. See appendix (2D joint prior) for further discussion.

4.3. Synthesizing Plausible Grasps

Setup and Evaluation Metrics. We evaluate human grasp synthesis on two datasets [16, 18]. HO3D is a real-world HOI dataset whose objects come from the YCB dataset [4], which has appeared in our training data. To test the generalization ability to novel objects, we also evaluate on a subset of 3D Warehouse used in Hasson *et al.* [18] (3DW). It is a synthetic dataset that our prior has never seen in training. Following prior work [22, 24], we evaluate grasp quality by 1) the amount of intersection between hands and objects (mean volume, maximum and mean depth), 2) the displacement of objects when placed into simulation [9], and 3) the contact hand region (ratio and area, where ratio is the percentage of grasps that have non-zero contact area). There is a trade-off between contact/simulation displacement and intersection. While the metrics characterize the grasp quality, no single metric alone is conclusive on grasp synthesis. So we also conducted a user study. We show users two human grasps randomly chosen from two methods and ask them to select their preferred one. We collected 440 and 380 answers from 22/19 users on HO3D and 3DW accordingly.

Table 2. **Comparison with Baselines:** We compare our synthesised human grasps against GraspTTA [22] and annotated grasps provided by datasets (GT) on HO3D and 3DW. We report table the intersection between meshes, displacement distance in simulation, and hand contact ratio and area (top). We also report preference percentages from users for pairwise method comparison on HO3D and 3DW (bottom).

		Intersection			Disp.	Contact	
		maxD↓	avgD↓	vol↓	avg ↓	ratio↑	area↑
HO3D	GT	**1.32**	0.37	6.16	2.32	0.95	0.15
	GraspTTA	2.44	0.61	**5.25**	2.89	1.00	0.23
	G-HOP	1.84	**0.31**	11.46	**0.95**	**1.00**	**0.23**
3DW	GT*	0.98	0.74	**1.70**	1.57	1.00	0.12
	GraspTTA	0.87	0.58	5.56	**1.54**	**1.00**	0.18
	G-HOP	**0.74**	**0.51**	17.40	1.85	0.93	**0.25**



Baselines. We compare with baseline GraspTTA [22] which is trained on in-domain data (3DW with annotated grasps). It learns to generate contact maps on hand and object which are then optimized along with hand pose be self-consistent during test time. We also compare with ground truth annotation in both datasets. While Grasping Fields [24] is also a representative method for grasp generation, their evaluation setup assumes a known object pose

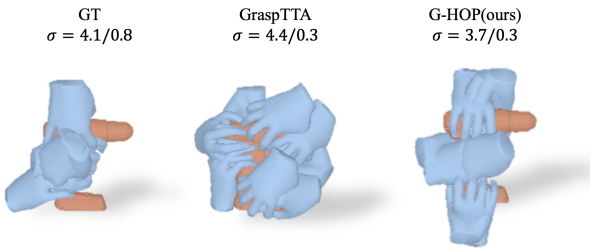


Figure 11. **Grasp Diversity:** 10 random grasps of a power drill. Although GraspTTA generates more diverse grasps, some of them are not plausible as they disregard object functions.

relative to the hand unlike ours, and randomizing this relative pose significantly affects their performance. We detail this further and report our results under their evaluation setting in supplementary.

Results. Fig. 9 visualizes two human grasp synthesis from each method for a given object. Annotated grasps (GT) in two datasets display different grasping styles. Semi-automatically generated grasps [18] sometimes do not look natural and tend to "over-grasp" as they are generated to maximize stability. GraspTTA is trained on the same dataset and shows similar over-grasp behavior while our grasps appear more natural. In contrast, G-HOP grasps objects from different directions while all of the synthesized hands make contact with the objects.

Grasp Diversity. We calculate the mean of standard deviations of hand vertices σ from 100 generations per object in the object/hand-centric frame on HO3D in Fig. 11. All methods show comparable diversity in the object-centric frame but both methods can improve on the diversity of finger articulation. Note that standard deviation on its own is not a good metric as diverse samples may be implausible or ignore object affordance as visualized.

Grasp Characteristic. Fig. 10 visualizes the overall contact probability on hand across all generated grasps. The contact region of GraspTTA is centered at fingertips and (implausibly) even at the nail region shown on the back of the hand. Contact regions from G-HOP are distributed on both fingers and palm, which is more consistent with how humans use their hands [2].

Tab. 2 also reflects the same characteristics. Although G-HOP has higher intersection volume, it has lowest average intersection depth and largest contact area. It also achieves the best performance in terms of grasp stability on HO3D and comparable results on out-of-domain 3DW objects. In user studies, G-HOP is preferred against all methods on both datasets, even when comparing with ground-truth.

Ranking Grasps. Finally, we show that the proposed grasp score yields meaningful grasp ranking. In Fig. 12, we visualize top 2 and bottom 2 grasps out of 100 generations from G-HOP, evaluated by the proposed evaluation method. The ranking matches human's common sense. For

Table 3. **Ranking Grasps:** plausibility on HO3D over all grasps, along with the top and bottom 10% grasps ranked by G-HOP.

	maxD↓	avgD↓	vol↓	disp ↓	ratio↑	area↑
G-HOP (top 10%)	**1.74**	**0.31**	**10.57**	**0.71**	1.00	0.22
G-HOP (all)	1.84	0.31	11.46	0.95	1.00	0.23
G-HOP (bottom 10%)	1.87	0.33	13.11	1.41	1.00	0.23

Figure 12. **Ranking Grasps:** We visualize grasps with two highest scores (top) and two lowest scores (bottom) among 100 generated grasps from G-HOP.

example, power drills are often held in the middle; narrow side of bottles is often held upwards. Physically infeasible grasps are ranked low such as hands penetrating the mug. Furthermore, the worst two grasps out of 100 are still reasonable in most cases. Note that all the grasps we show to users are randomly chosen for fair comparison. Quantitatively, top-ranked grasps in Tab. 3 show reduced simulation displacement and less intersection, validating our ranking approach's efficacy.

5. Conclusion

In this work, we propose a method to jointly generate 3D shape of HOI given an object category. Our method is the first to generate HOI across such diverse categories. The learned prior G-HOP can serve as generic prior for relevant tasks like reconstructing interaction clips and human grasp synthesis, and we find that it leads to better performance than current task-specific baselines. Despite the encouraging results, we are aware of several limitations: current method requires category information as input which may prevent the model from further scaling up; there is no explicit mechanism to guarantee contact; and the model is still not at a scale comparable to generative models in other domains due to limited training data. Nevertheless, we believe that our work takes an encouraging step towards scaling up a general understanding of hand-object interactions.

Acknowledgements. The authors would like to thank Hanwen Jiang and Korrawe Karunratanakul for clarifying baselines. We also thank Fu-Jen Chu and Ruihan Gao for their detailed feedback on the manuscript. Yufei's PhD research is partially supported by a Google Gift.

References

[1] Dafni Antotsiou, Guillermo Garcia-Hernando, and Tae-Kyun Kim. Task-oriented hand motion retargeting for dexterous manipulation imitation. In *ECCV Workshops*, 2018. 2

[2] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *IROS*, 2019. 2, 8

[3] 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 *ECCV*, 2020. 2, 5

[4] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *ICAR*, 2015. 7

[5] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021. 5

[6] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *ECCV*, 2022. 2

[7] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *CVPR*, 2023. 2

[8] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020. 5

[9] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021. 7

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 6

[11] Congyue Deng, Chiyu Max Jiang, C. Qi, Xinchen Yan, Yin Zhou, Leonidas J. Guibas, and Drago Anguelov. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. *CVPR*, 2023. 2

[12] George ElKoura and Karan Singh. Handrix: animating the human hand. In *SCA*, 2003. 2

[13] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, 2018. 2

[14] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018. 2

[15] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. Contactopt: Optimizing contact to improve grasps. *CVPR*, 2021. 2

[16] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 2, 6, 7

[17] Shreyas Hampali, Tomás Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-hand 3d object scanning from an rgb sequence. *CVPR*, 2023. 2

[18] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2, 7, 8

[19] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 5

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[21] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia*, 2022. 2, 6

[22] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, 2021. 2, 7

[23] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv*, 2023. 2

[24] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*, 2020. 2, 7

[25] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *ICCV*, 2023. 2, 3

[26] Jun-Sik Kim and Jung-Min Park. Physics-based hand interaction with virtual objects. In *ICRA*, 2015. 2

[27] 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

[28] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *CVPR*, 2023. 2

[29] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *CVPR*, 2023. 2

[30] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *CVPR*, 2022. 2

[31] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[32] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, 2022. 5

[33] Jeffrey Mahler, Matthew Matl, Vishal Satish, Michael Danielczuk, Bill DeRose, Stephen McKinley, and Ken Goldberg. Learning ambidextrous robot grasping policies. *Science Robotics*, 2019. 2

[34] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360° reconstruction of any object from a single image. In *CVPR*, 2023. 2

[35] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *ICCV*, 2019. 2

[36] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR*, 2022. 2, 3

[37] Aditya Prakash, Matthew Chang, Matthew Jin, and Saurabh Gupta. Learning hand-held object reconstruction from in-the-wild videos. *arXiv preprint arXiv:2305.03036*, 2023. 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[41] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 2017. 3

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2

[43] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *ICLR*, 2023. 2

[44] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 5

[45] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, 2019. 2

[46] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 2, 3

[47] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017. 3

[48] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023. 2, 3

[49] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *CVPR*, 2023. 2

[50] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*, 2022. 5

[51] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022. 2, 3, 6

[52] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023. 2, 4, 5, 6

[53] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 2

[54] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 2