GraspDiffusion: Synthesizing Realistic Whole-body Hand-Object Interaction

Patrick Kwon*
University of Central Florida
yo564250@ucf.edu

Chen Chen
University of Central Florida
chen.chen@crcv.ucf.edu

Hanbyul Joo
Seoul National University
hbjoo@snu.ac.kr

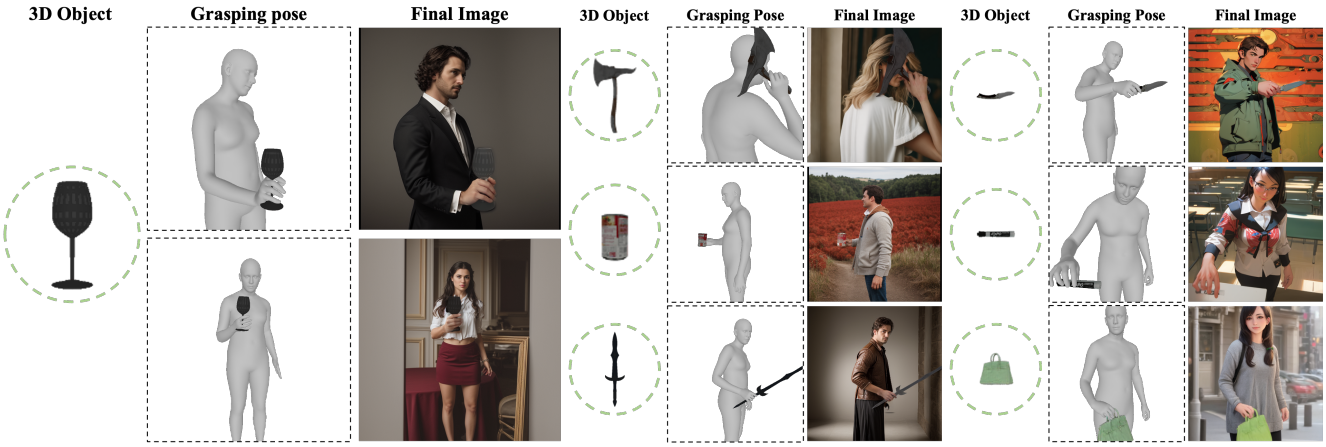


Figure 1. Given an object mesh and its relative position, GraspDiffusion generates whole body grasping 3D poses, which is subsequently used as guidance for creating human-object interaction scenes. As shown, GraspDiffusion can synthesize images with valid human-object interactions for various types of objects. Note that the bottom-right sample (a green bag) was created from an object image, which was made into a 3D using TripoSR [75], further paving the way for various use cases.

Abstract

Recent generative models can synthesize high-quality images, but they often fail to generate humans interacting with objects using their hands. This arises mostly from the model's misunderstanding of such interactions and the hardships of synthesizing intricate regions of the body. In this paper, we propose ***GraspDiffusion****, a novel generative method that creates realistic scenes of human-object interaction. Given a 3D object, GraspDiffusion constructs whole-body poses with control over the object's location relative to the human body, which is achieved by separately leveraging the generative priors for body and hand poses, optimizing them into a joint grasping pose. This pose guides the image synthesis to correctly reflect the intended interaction, creating realistic and diverse human-object interaction scenes. We demonstrate that GraspDiffusion can successfully tackle the relatively uninvestigated problem of generating full-bodied human-object interactions while outperforming previous methods. Our project page is available at* https://yj7082126.github.io/graspdiffusion/

*Work done while at Naver Webtoon.

1. Introduction

The recent advent of diffusion-based generative models [28, 70, 71] has demonstrated significant success in producing high-quality visual content [59, 62, 65, 66, 69]. When trained on large datasets, these models can coherently synthesize images of various subjects corresponding to given textual/visual cues. However, despite their strong performance, generative models struggle to comprehend and visualize everyday hand-object interactions. This limitation hinders their broader adoption in generative model-based content creation pipelines.

One challenge in generating images with high-quality hands arises from the fact that hands occupy only marginal areas within a full-bodied human image, yet have a complex anatomical structure that presents a wide variety of possible hand poses. Hands come in varying shapes, sizes, orientations, and multiple finger joints that can bend in various degree to support diverse hand poses. Moreover, hands are usually interacting with various objects, making the distribution of hands highly complex and convoluted, leading to faulty generation results (e.g., distorted hand poses, multiple arms from a shoulder, uncanny hand shapes). Examples

of such inaccurate generation are displayed in Fig. 2.

While several papers [53, 61, 77] have applied inpainting for hand region refinement, they only focus on situations where the hand is not interacting with other objects, making it impractical for most use cases. In order to generate realistic grasping hands for a given object, the models must understand the semantics and functionalities provided by the object–a concept well known as affordance [19]. While Affordance Diffusion [86] and HOIDiffusion [89] creates an image of a single grasping hand using affordance, these methods only represent the explicit physical contact between a human hand and the object (devoid of any human identity), and fail to convey the spatial / orientational non-contact relationships, making them unsuitable for understanding human affordance. For instance, when using a cell phone, the relationships between the human's face and torso with respect to the phone should be considered as part of interaction along with the hand touching the phone. This requirement necessitates the development of an pipeline that creates identifiable human-object images where the human's body is visible, such that the implicit, non-contact relationships are well captured.

In this paper, we present GraspDiffusion, a novel method for generating interaction images with realistic hands and a clear human identity from a single 3D object input. Instead of relying on textual prompts, which tend to be ambiguous in describing complex interactions [57], we first utilize a diffusion model to generate full-bodied grasping poses [60, 67] conditioned on the object and its position. The generated 3D pose parameters are provided as conditions for the next stage to create high-quality image samples. Specifically, we train multiple conditional models [56, 88] and a novel cross-attention modulation scheme [4, 15] to correctly convey the interaction context without harming the diversity of the generated samples. To overcome the lack of a large-scale image-3D pose paired dataset for human object interaction, we also propose a dataset annotation scheme to gather 3D annotations from previous image-based interaction datasets [11, 20, 21, 32, 48].

As far as our knowledge, our pipeline is the first approach to generate full-bodied HOI images from a given object information, such that both explicit and implicit interactions are portrayed in a physically plausible manner. Our experiments on several metrics show that GraspDiffusion outperforms similar approaches in generating high-quality images with realistic interactions and valid 3D grasping poses. We also display cases where GraspDiffusion can be conditioned with a single object image and support various artistic styles, proving its efficacy as a practical solution for AI practitioners in content creation. The contributions of this work are summarized as follows.

- We propose GraspDiffusion, a novel generative pipeline that synthesizes realistic, full-bodied HOI images.

Figure 2. Comparison between our method and previous approaches on generating HOI images. While previous methods can generate images conditioned on human pose and refine hand shapes, they are prone to erroneous object creation (top row) or faulty interaction synthesis (bottom row).

- We divide the pipeline into two stages to facilitate both physically plausible body poses along with identity and style diversity. The first stage provides a rich 3D prior to create lifelike interaction poses, which are then provided as 3D conditions for the image generation stage which prodces images with high quality and diversity.
- We devise an annotation pipeline for an image-3D paired grasping dataset to facilitate the training of HOI image generation.
- Our experiments in generating realistic images and their paired 3D grasping poses demonstrate the efficacy and substantial performance improvements over previous SOTA methods, providing advanced physical pose plausibility and perceptive visual quality.

2. Related Work

Conditional Image Generation. To provide additional fine-grained, spatial conditions for diffusion models, ControlNet [88] and T2I-Adapter [56] proposed using image-level signals to control the generation process, which includes using 2D human keypoint skeletons [10, 84] for human pose guidance and depth maps for better depth perception. While the improvements in conditional image generation have been significant, synthesizing humans interacting with objects hasn't reached the same level of improvement.

Several papers [53, 61, 77] proposed to refine the hands of images generated using Controlnet-based inpainting, but

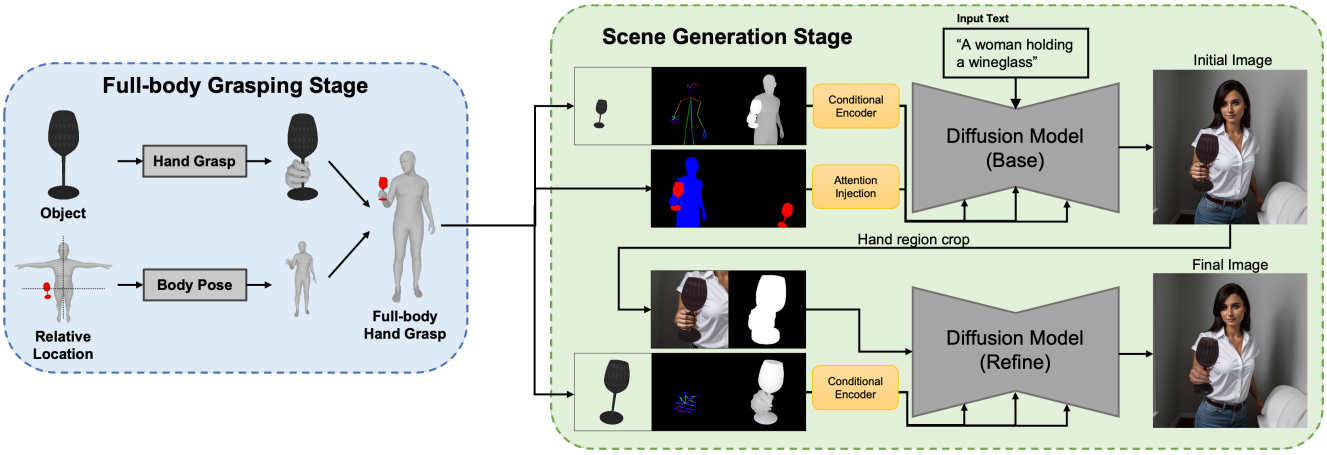


Figure 3. We present a two-stage pipeline to generate realistic human-object-interaction images. The first stage takes a single object model and its human-centric location to synthesize a 3D full-bodied grasping pose, providing scene-level context for image generation. The second stage takes reference from the 3D grasping pose, conditionally generating high-quality images.

does not count as a direct solution to human-object interaction. Others focused on identifying contacts for inpainting a new hand or object for a given scene [41, 55, 83, 86], yet they are limited in camera views and require a prior scene context, making them infeasible in practical scenarios. HOIDiffusion [89] also generates images from synthesized grasps, but is limited to hand-centric views. Compared to HanDiffuser [57], which applied the injection of hand embeddings during image generation to create realistic hands, our pipeline focuses more on the joint synthesis of hand and object, and uses spatial guidance to better direct the generation process towards a 3D scene context.

Grasp Synthesis. Synthesizing a hand grasp consisting of a given object and a hand model is important in understanding human-object interaction, and is a widely studied task in robotics, graphics, and computer vision. While the focus in robotics is to make stable grasps for a given object in simulation / real life [17, 46, 76, 78, 79], in computer vision and graphics the focus is to make plausible grasps that are physically plausible, and generate grasps for either hands [13, 14, 30, 34, 47, 50, 90, 92] or a human [9, 18, 54, 73, 74, 81, 91].

Thanks to the advent of human-object interaction datasets [5, 8, 12, 16, 22, 24, 42, 51, 72], many grasp synthesis methods achieved high performance in generating plausible grasps. Yet most existing datasets have issues with data scalability and variability, especially when it comes to color image-3D paired datasets. While hand-object interaction datasets like DexYCB [12], ARCTIC [16] and HOI4D [51] consist of richly annotated image data, they are rather focused only on the hand and object recorded in an egocentric manner, devoid of any human identity and spatial / orientational non-contact relationships. Although BEHAVE [5] contains both RGB video sequences with 3D annotation, the image quality and motion sensors worn by the subject make it difficult to use as a realistic image dataset. To overcome this issue, we used traditional human-object interaction datasets [11, 20, 21, 32, 48] along with annotation tools [35, 39, 43, 44, 49] to create pseudo-3D annotations for the 2D image.

Building our insight from similar approaches, [74, 91], we focus on leveraging priors from a full-body pose model and a hand-grasping model. Compared to previous approaches [9, 18, 73, 81], instead of generating a motion sequence, we synthesize the pose parameters for the 3D parametric models [60, 67] using a diffusion model.

3. GraspDiffusion

Fig. 3 illustrates the proposed architecture. Starting with a 3D object mesh and its position within the human-centric coordinate system (originating at the pelvis joint), GraspDiffusion synthesizes realistic images portraying a human interacting with the object, with a significant portion of the human body visible to be considered "full-body". In the initial full-body grasping stage, we generate the pose parameters for the human body model [60] interacting with the 3D object mesh (Section. 3.2). In the scene generation stage, we extract geometric structures from the pose parameters to guide the generation of realistic images, leveraging a latent diffusion model [66] along with spatial encoders and a cross-attention modulation scheme (Section. 3.3).

3.1. Preliminaries

Diffusion Models. Diffusion models [28, 70] are a group of generative models that interpret the data distribution $p(x)$ as a sequential transformation from a tractable prior distribution $p(x_T) \sim \mathcal{N}(0, I)$. During training, the model uses a forward noise process $q(x_t|x_{t-1})$ that gradually adds a

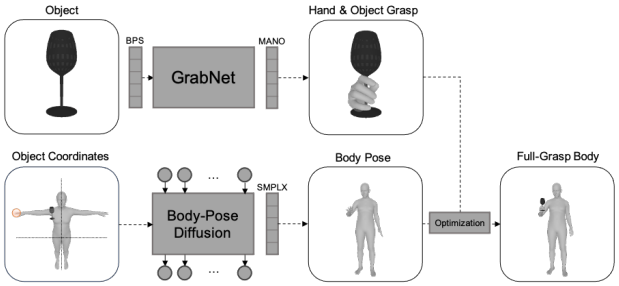


Figure 4. Full-body grasping pipeline. We separately leverage a hand-grasping model [72] and a body-pose diffusion model, and perform a joint optimization into a full-bodied grasping pose.

small amount of noise to a clean data sample x_0 towards $p(x_T)$. At the same time, the model learns a backward noise process $p(x_{t-1}|x_t)$ implemented as a neural network, which is trained to remove the noise from the before generating samples from $p(x_T)$. For latent diffusion models [66], the diffusion process is performed in the latent space of a trained autoencoder model [38], guided by a conditional text embedding derived from the CLIP [64] mechanism.

3D Parametric Models. For the hand model, we use the MANO [67] differentiable model, in which we input the full finger articulated hand pose $\theta_{\text{h}} \in \mathbb{R}^{15\times3}$, wrist translation $t_{\text{h}} \in \mathbb{R}^3$ and global orientation $R_{\text{h}} \in \mathbb{R}^3$ and get a 3D mesh $\mathcal{M}_{\text{h}}$ with vertices $\mathcal{V}_{\text{h}}$. For the full-body model, we use the SMPL-X [60] differentiable model, in which we input the full-body pose $\theta_{\text{b}} \in \mathbb{R}^{21\times3}$, the full finger articulated hand pose $\theta_{\text{h}} \in \mathbb{R}^{15\times3}$ for both hands, the root translation $t_{\text{b}} \in \mathbb{R}^3$ and global orientation $R_{\text{b}} \in \mathbb{R}^3$ and get a 3D mesh $\mathcal{M}_{\text{body}}$ with vertices $\mathcal{V}_{\text{b}}$.

3.2. Full-Body Grasping Pipeline

Building on prior approaches [74, 91], we separately generate hand grasps and body poses in creating a whole-body grasping pose. Specifically, we take a 3D object mesh, its relative location to the human root, and the contacting hand orientation (left or right) as the input, to generate an SMPL-X mesh that grasps the given 3D object with a realistic body pose and hand-object contact.

The input object mesh is used to generate a plausible MANO [67] hand grasp, for which we utilized GrabNet [72], a conditional variational autoencoder (cVAE) that produces hand grasps conditioned on the Basis Point Set (BPS) [63] of the given object. Separately trained for left and right-hand grasps, GrabNet generates MANO parameters $(\theta_{\text{h}}, t_{\text{h}}, R_{\text{h}})$ that grasps the given object, displaying accurate contact and high generalization for unseen objects.

The object's relative location $t_{\text{obj}} \in \mathbb{R}^3$ and the hand orientation $c_{\text{left}}, c_{\text{right}} \in \{0, 1\}$ is then used to create a body pose that not only roughly positions its hand in the desired object location, but also reflects the appropriate implicit relationships required for a plausible grasping body pose [36, 73]; whether the head is correctly oriented towards the object, the arms are correctly extended and the torso is leaning towards the object. To achieve this, we utilize a diffusion generative model trained to generate SMPL-X pose parameters $(\theta_{\text{body}}, R_{\text{body}})$ conditioned on an object location and whether to use the right/left hand for contact. The loss is defined as

$$\mathcal{L}_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,I),t}[||\epsilon - \epsilon_\theta(x_t, t, c)_2^2||], \tag{1}$$

where $c = [t_{\text{obj}}, c_{\text{left}}, c_{\text{right}}] \in \mathbb{R}^5$ and $x \in \mathbb{R}^{132}$, which consists of the 6 DoF global orientation and body pose.

We then apply the finger articulation of the hand grasp to the body pose, creating an initial full-body grasping pose. To correctly align the 3D hand-object grasp with the human body, we optimize over the rotation (R_{h}) and translation (t_{h}) of the MANO hand model while retaining the original finger articulation. Focusing on the palm region of the hands, given vertices $\mathcal{V}_h^p$ (output palm vertices from MANO) and corresponding vertices $\mathcal{V}_b^p$ (output palm vertices from SMPL-X), we align them using:

$$E(R_{\text{h}}, t_{\text{h}}) = \frac{1}{|\mathcal{V}_h^p|}\sum_{i=1}^{|\mathcal{V}_h^p|} d_{\text{vv}}(\mathcal{V}_{h_i}^p, \mathcal{V}_{b_i}^p), \tag{2}$$

where d_{vv} represents the L^1 distance between the two vertices in the 3D space. The optimized $(R_{\text{h}}, t_{\text{h}})$ is used to transform the 3D object, correctly positioning it within the full-body grasping pose as it was for the hand-object grasp, completing the grasping pose.

3.3. Scene Generation Pipeline

Given the 3D body pose pose, the scene generation pipeline extracts multiple spatial conditions as conditions to a pre-trained [66] model to create consistent images of human-object interaction. Optionally, the pipeline can refine the image focused on the hand-object region to further adjust interaction and correct erroneous details.

To precisely control the human-object image's generation, we render three spatial conditions from the full-body grasping output. We first render the skeleton (s^i) of the SMPL-X body, consisting of body and hand joints, to ensure realistic human proportions within the generated image. We also use the joint depth map (d^i) from the SMPL-X and object model to provide depth information. Lastly, we render the occluded object with ambient lighting (o^i) to preserve its appearance while relighting it. To apply conditions, we chose the CoAdapter [56] approach, which allows flexibility in handling multiple conditions. For each condition, we separately apply an adapter $\mathcal{F}_{\text{AD}}$ and perform a weighted sum to create feature $\mathbf{F}_c$, which can be written as

$$\mathbf{F}_c = \sum_{k\in\{s,d,o\}} \omega_k \mathcal{F}_{\text{AD}}^k(k^i). \tag{3}$$

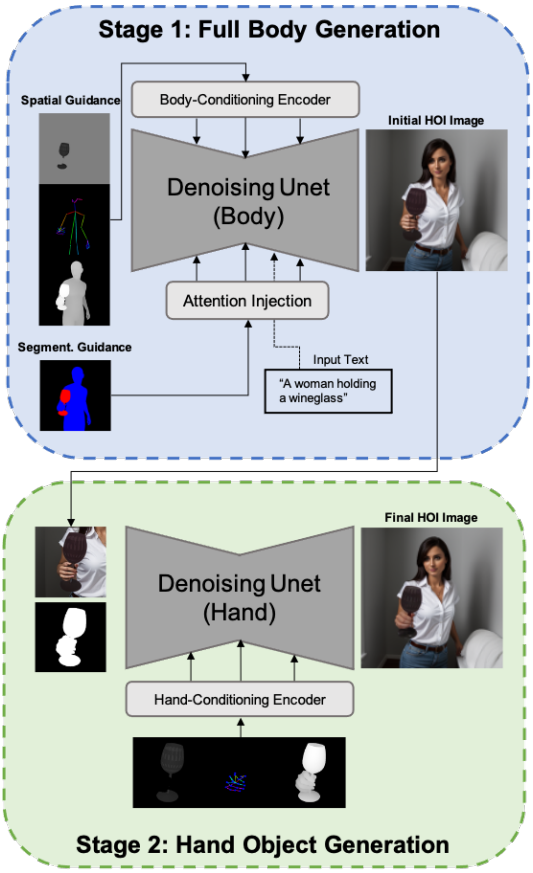


Figure 5. Scene generation stage. We inject three image conditions and semantic segmentation images as guidance for the generation of a high-quality HOI image. We then use the same types of renderings centered on the hand-object region to refine the hand quality.

During training of the conditioning pipeline, we fix the parameters of the baseline U-Net model and only optimize the conditional adapters, reducing the risk of the model converging to the dataset's style. This decision allows us to control the image style by applying LoRAs [29] or using finetuned Stable Diffusion models during inference time, making it suitable for real-world application tasks requiring personalized image generation.

$$\mathcal{L}_{ADM} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,I),t,\mathbf{F}_c}[||\epsilon - \epsilon_\theta(z_t, t, c_{\text{text}}, \mathbf{F}_c)_2^2||]. \quad (4)$$

For hand-object refinement, we utilize the full-body grasping output to produce a joint hand-object mask and spatial conditions: hand skeleton information (s_h^i), a joint hand-object depth map (d_h^i), and the occluded rendered object (o_h^i). These masks and conditions serve as inputs to the hand refinement adapters, which is akin in structure to the body conditioning adapters yet trained separately. This refines the structure and appearance of both hand and object while preserving visual integrity. A full illustration of the procedure is shown in Fig. 5.

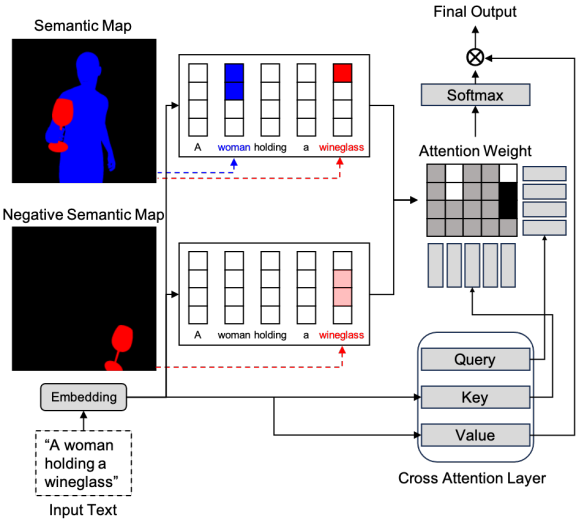


Figure 6. Attention Injection Scheme. During inference, we inject the human/object semantic maps into the cross-attention layers as guidance, encouraging the generation process to be focused on the segmented regions. We also apply a negative semantic map for the object to avoid undesired cases where the opposite hand interacts with the object.

To address the issue of erroneous interactions, in which interactions may occur from locations other than the intended area (examples on Fig. 8), we introduce a training-free guidance method motivated from prior zero-shot semantic image synthesis techniques [4, 15]. We first render binary segmentation masks from the posed human and object 3D model (m^i, m_o^i), which are then sent to the cross-attention layers as guidance, down-sampled to match the resolution of each layer. Specifically, we create an input attention matrix $A \in \mathbb{R}^{N_i \times N_t}$ from the masks, applied to the cross-attention layers to encourage attention towards the intended region. We also modify the original procedure through the usage of a negative mask; specifically, we create a pseudo object segmentation map m_{no}^i which, instead of using the intended hand, is using the opposite hand to grasp the 3D object model. This segmentation mask is then subtracted from the input attention matrix, disencouraging the generation in unintended locations.

4. Experiments

4.1. Dataset Construction

To compensate for the lack of realistic, 3D-annotated human-object interaction datasets, we designed an annotation pipeline through which we leveraged previous interaction datasets [11, 20, 21, 32, 48] to construct a pseudo-3D interaction dataset. Specifically, we utilized the human-object interaction images from HICO-DET [11] and V-COCO [21], which contain a large variety of possible in-

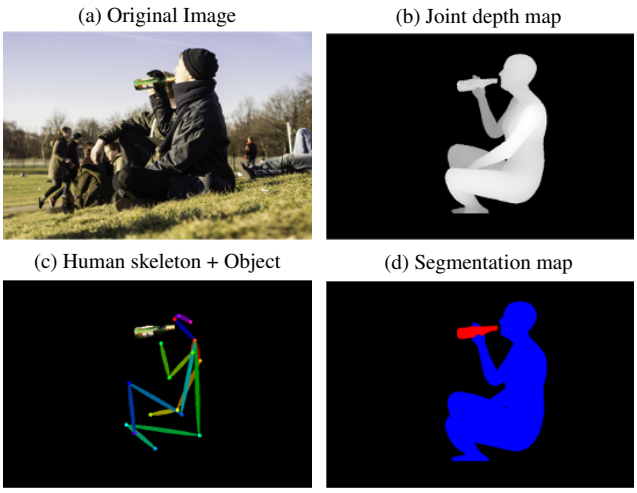


Figure 7. Processed HICO-DET [11] dataset sample. We jointly extract the joint depth map for the object and human 3D pose, the human skeleton joint map, and semantic segmentation maps.

teractions and annotations for the human body and object type.

For both datasets, we first filtered the images so that each image included at least one visible human with a reasonable screen size, along with at least one identifiable hand. For datasets like HICO-DET, multiple humans within an image might be interacting with multiple objects, making it necessary to correctly identify the main object of interaction. We use the BLIP-2 language model [45] to perform a Visual Question Answering task, which outputs an object type from the input image. Using the object text, we use GroundingDINO [49] to get the object location and detect the segmentation map of the object, and employ a state-of-the-art depth estimation model [35] to estimate the depth map of the object. Note that for V-COCO, we use the original annotated information for objects Meanwhile, we also estimate the 3D SMPL-X parameters for the human, using the annotated bounding box and HybrIK [43, 44] 3D human pose and shape estimator. For the human skeleton, we use the annotated Halpe dataset [32] for the HICO-DET dataset and the DWPose estimator [10] for the V-COCO dataset. Among the processed images, we identify images with sufficiently large hands portrayed and reserve them for the hand-object refinement model training. To estimate the 3D MANO parameters, we use the ACR [87] hand pose and shape estimator. To further augment the dataset, we use the BEHAVE interaction dataset [5] that comes with SMPL-X parameters and object 3D models, and processed it in a similar manner to create joint image interaction pairs (image, text, depth map, skeleton, segmentation).

For the hand refinement modules, we process the DexYCB dataset [12], the RHD dataset [94] that also comes with 3D annotations and video data. By processing the

Table 1. Quantitative comparison on full-bodied generation

Methods	FID ↓	KID ↓	CLIPScore ↑
LDM (finetuned) [66]	41.23	1.45×10^{-2}	0.671
ControlNet [88]	32.76	1.23×10^{-2}	0.71
Champ [93]	40.63	2.23×10^{-2}	0.739
Ours (w/o atttention)	**22.55**	5.63×10^{-3}	0.717
Ours	22.88	$\mathbf{5.55 \times 10^{-3}}$	**0.767**

Table 2. Quantitative comparison on hand-object generation

Methods	FID ↓	KID ↓	Hand Contact ↑
ControlNet [88]	99.38	7.70×10^{-2}	58.17
HandRefiner [53]	92.48	7.11×10^{-2}	61.45
Affordance Diffusion [86]	-	-	65.69
Ours	**64.67**	$\mathbf{4.36 \times 10^{-2}}$	**97.94**

MANO parameters [67] and the object 3D models, we also acquired image interaction pairs specifically focused on the hand region. To broaden the hand-object interaction distribution, we include a subset of the preprocessed HICO-DET dataset, cropped on the hand-object bounding box, as additional training data for hand refinement. In total, we collected 25K joint interaction pairs of the full human-object interaction scenes to train the scene generation pipeline, and 15K joint interaction pairs of the hand-object interaction scenes to train the hand refinement pipeline.

4.2. Implementation Details

For the full-body grasping pipeline, we train the body-pose diffusion model on the GRAB [72] dataset, which captures the full-bodied 3D SMPL-X interaction sequences with various objects. After downsampling the motion sequences of GRAB to 30 fps, we collect all frames that has more than 40 contacting vertices between the object and the subject's right hand, to collect feasible grasps for the training. We then adopt the Adam optimizer [37] with a learning rate of 5×10^{-4}. We train the model with batch size of 2,048 for 50k steps, using 2 RTX 6000 GPUS. For the diffusion schedule, we adapt a cosine noise schedule with $T = 1000$.

To implement the scene generation pipeline, we train the two modules with the aforementioned custom datasets, using 20K of the interaction pairs. Employing the Stable Diffusion v1.5 [66] as a base model with parameters frozen, we train the conditional modules with a constant learning rate of 10^{-4}, for 200 epochs on four A100 GPUS which costs approximately 28 hours. For inference, we used a linear multistep scheduler [33] with 30 inference steps using a classifier-free guidance [27] of 3.5. We also support inference using personalized Stable Diffusion models other than the Stable Diffusion v1.5 model used during training, and display results with different models in Fig. 9.

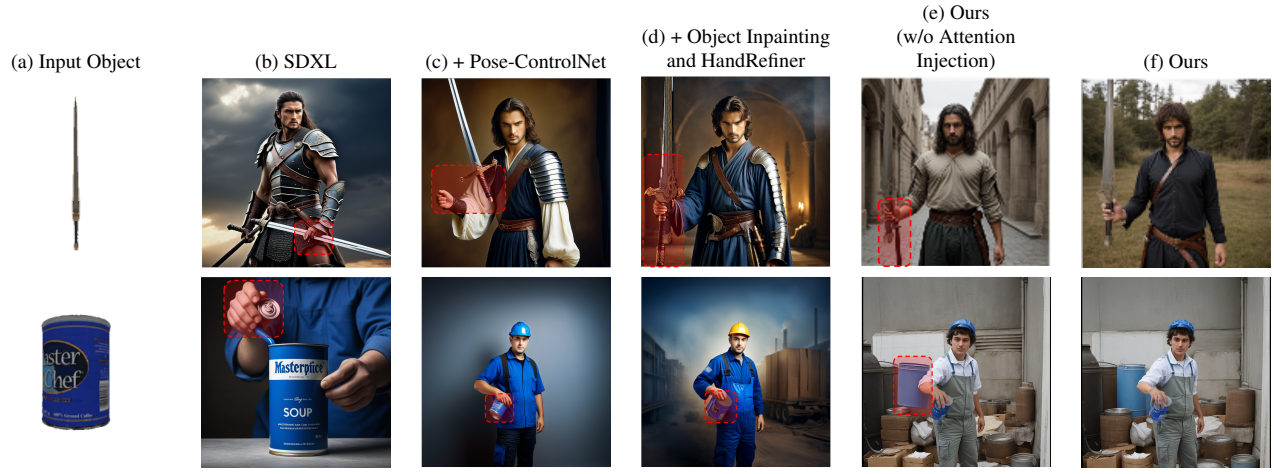


Figure 8. Qualitative results. We compare HOI images generated by different methods based on a input object (first column). Note that except for the second column, all images were based on the same human pose and object location created from our grasping pipeline. While other methods display erroneous interactions (e.g. multiple objects, object appearance distorted, physically implausible interactions, color blending), which are marked with red segments, our scene-generation pipeline can correctly convey the interaction intention from the full-body grasping pipeline.

Table 3. Grasping pose evaluation

Methods	Contact ratio ↑	Pose Valid Error ↓	Displacement ↓
GOAL [73]	0.461	0.504	6.135
FLEX [74]	0.540	0.252	7.794
COOP [91]	0.841	0.239	4.679
Ours	**0.909**	**0.111**	**2.696**

4.3. Quantitative Results

Full-body generation quality To assess the generation quality, we adopt Frechet Inception Distance (FID) [26] and Kernel Inception Distance (KID) [6]. We compare our results with three baseline models: (1) a finetuned Stable Diffusion v1.5 model [66] conditioned only by the text description; (2) ControlNet [88] with multiple control input; (3) Champ [93], a human image animation method that uses SMPL-X sequences. For Champ, we separately generate a human image as reference and control the body pose of the reference image using Champ's guidance encoders, without using its motion module. We used the 5K testing set from our novel human-object dataset for comparison. In addition, to assess the alignment between the intended text prompt and the generated image's interaction context, we use CLIPScore [25] as an additional evaluation metric. As shown in Table 1, our method can improve image quality and prompt alignment in generating images with human-object interaction.

Hand-grasp generation quality We also assess the quality of hand-centric images from the hand refinement pipeline, based on both image quality and plausible hand-object pose. To measure instances of successful hand-object contact, we adopt the contact evaluation setup in Affordance Diffusion [86] and utilize a widely used hand-object detector [68] to measure the object's contact status. We compare our results with three baseline models : (1) a depth-based ControlNet, (2) HandRefiner [53], and (3) Affordance Diffusion [86]. We evaluated on a subset of the DexYCB dataset [12]. We report the results in Table 2, which shows that our method is capable of outperforming previous methods in creating hand-object images with accurate contact.

3D pose evaluation To evaluate the plausibility of grasping poses for different objects and positions, we constructed a test set of unseen objects distributed far from the original range of the training dataset. We choose 10 novel 3D objects from Dex-YCB [12] and 10 human body shapes, and for each pair, we position the object at 64 random 3D positions, relative to the human body's pelvis joint position. Specifically, the x-coordinate (the horizontal position in our paper) ranges from -0.5m to 0.5m, the y-coordinate (the vertical position in our paper) ranges from -0.8m to 0.8m, and the z-coordinate (the direction where the human model is facing) ranges from 0.0m to 0.8m, with the pelvis joint position as its origin.

For pose evaluation, we utilize VPoser [60] to measure the L2 loss of vertex reconstruction from the body poses as a pose-valid error, given that an implausible body pose will result in a higher pose-valid error. We compare our approach with two prior methods that are trained on the GRAB dataset [72] and support generating full-body grasps for different object translations; GOAL [73] and COOP [91]. For GOAL, we only evaluate the grasping pose generation with optimization (GNet) and set the x-coordinate of the object translation to 0 due to the fact that GOAL does not work when the objects are out of distribution in the horizontal plane [74]. We present the results in Table 3. The results demonstrate that our model is capable of generating

Table 4. Ablation studies on architecture choice

Methods	FID ↓
Ours	**22.88**
w/o object rendering	29.53
w/o human skeleton	26.35
w/o joint depth	24.37

authentic grasping poses for objects in various positions.

Ablation Studies During the scene generation pipeline, we utilize different structural renderings from the 3D grasping model as conditions to generate a realistic image; the object rendering defines the appearance and location of the target object, the skeleton map gives a precise human pose depiction, and the depth map maintains geometric consistency. To investigate the importance of each factor, we measure the FID scores for cases where only two of the three conditional modules are provided during inference. As presented in Table 4, our full model setting outperforms other settings with missing conditions.

In Table 1, we also display results for the main setting without attention injection, which alleviates the risk of generating erroneous interactions. While the setting without attention injection has a slightly better FID / KID score, our full setting shows a substantial increase in CLIPScore, signifying that the generated images successfully adhere to the given interaction context.

User Study We conducted a user study to measure the perceptual quality and geometric consistency of our pipeline. We asked 28 participants to compare images that were generated based on the same 3D grasping pose and object, one generated using multiple ControlNets [88] and HandRefiner [53] and the other using GraspDiffusion. The participants were asked two types of questions: (a) which image is more realistic and plausible, and (b) given the original grasping information, which one follows faithfully to the grasping context. In total, 92.4% of the votes preferred our method over the baseline on plausibility, while 96.4% preferred based on following the given context.

4.4. Applications

By utilizing a 3D full-body grasping model generated from an object input, GraspDiffusion can provide a practical solution for AI practitioners who intend to use generative AI for their artwork such as advertisements, illustrations, and comic books. To alleviate the requirement of an object mesh model, our pipeline also supports using 3D reconstruction models [52, 75, 80] that can recover 3D mesh models from a single image input. Moreover, our pipeline can support various personalized image domains, including (but not limited to) realistic, anime, pixel art style, and more. In Figure 9 we present results from the same 3D grasping pose, using diverse personalized text-to-image models to support different art styles and backgrounds. Further examples are provided in the Supplementary Material.

Figure 9. Example results from different models. We display generation results from our pipeline with the same object and body pose, but with different personalized Stable Diffusion models that were acquired from CivitAI [1] and Huggingface [2].

5. Discussions and Limitations

While GraspDiffusion can produce humans with detailed finger articulation and accurate object interaction, several samples exhibit discrepancies between the body's skin texture and the refined hand's skin texture. We account this issue to the shortage of balanced, high-quality data samples during training, and opt to construct additional interaction samples to facilitate high-quality generation. In the future, we aspire to extend our pipeline toward scene-level generation that involves interaction between multiple humans and objects, and to better define interactions in the form of user-controllable text prompts. We also look forward to synthesize zero-shot interaction motions by leveraging image-to-video diffusion models [7, 23, 40, 85].

6. Conclusion

We present an image generation pipeline that is the first to explicitly target realistic human-object interaction. The resulting images exhibit both explicit (hand-object contact, realistic hand grasp) and implicit human-object interaction (human gaze, body direction), without requiring any auxiliary conditions other than a 3D object mesh and its relative position. The results demonstrate our method's effectiveness in creating images with plausible hand poses, while preserving the given object's identity. In the future, we plan to extend our pipeline towards generating various types of interaction (e.g. human-human interaction, specialized hand-object interaction), while further demonstrating the effectiveness of our pipeline in video generation and synthetic dataset creation for interaction detection.

References

[1] Civitai. https://civitai.com. Accessed: 2024-09-03. 8

[2] Huggingface. https://huggingface.co. Accessed: 2024-09-03. 8

[3] Sketchfab. https://sketchfab.com/. Accessed: 2024-09-03. 15

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 5, 14

[5] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15914–15925, 2022. 3, 6

[6] Mikolaj Binkowski, Danica J. Sutherland, Michal Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 7

[7] A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, abs/2311.15127, 2023. 8

[8] 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 *The European Conference on Computer Vision (ECCV)*, 2020. 3

[9] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. *2024 International Conference on 3D Vision (3DV)*, pages 464–473, 2023. 3

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:172–186, 2018. 2, 6

[11] Yu-Wei Chao, Yunfan Liu, Michael Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389, 2017. 2, 3, 5, 6

[12] 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. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9049, 2021. 3, 6, 7, 15

[13] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20545–20554, 2021. 3

[14] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomás Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. *arXiv preprint arXiv:2403.17827*, 2024. 3

[15] Guillaume Couairon, Marlene Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2174–2183, 2023. 2, 5

[16] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12943–12954, 2023. 3

[17] Haoran Geng and Yun Liu. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3868–3879, 2023. 3

[18] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and P. Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. *Computer Graphics Forum*, 42, 2022. 3

[19] James Jerome Gibson. The ecological approach to visual perception. *The Journal of Aesthetics and Art Criticism*, 39 (2):203–206, 1980. 2

[20] Georgia Gkioxari, Ross B. Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2017. 2, 3, 5

[21] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *ArXiv*, abs/1505.04474, 2015. 2, 3, 5

[22] Vladimir Guzov, Torsten Sattler, and Gerard Pons-Moll. Visually plausible human-object interaction capture from wearable sensors. *ArXiv*, abs/2205.02830, 2022. 3

[23] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 8

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

[25] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. 7

[26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, 2017. 7

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 6

[28] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances In Neural Information Processing Systems*, 2020. 1, 3

[29] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 5

[30] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11087–11096, 2021. 3

[31] Hanwen Jiang, Qixing Huang, and Georgios Pavlakos. Real3d: Scaling up large reconstruction models with real-world images. *arXiv preprint arXiv:2406.08479*, 2024. 13

[32] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. *ArXiv*, abs/2007.11858, 2020. 2, 3, 5, 6

[33] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022. 6

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

[35] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 6

[36] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. 2024. 4

[37] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[38] 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. 4

[39] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 3

[40] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jia-Liang Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fan Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Peng-Yu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhen Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Z. Xu, Yang-Dan Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *ArXiv*, abs/2412.03603, 2024. 8

[41] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A. Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17089–17099, 2023. 3

[42] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10118–10128, 2021. 3

[43] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 3, 6

[44] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023. 3, 6

[45] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 6

[46] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8068–8074, 2022. 3

[47] Peiming Li, Ziyi Wang, Mengyuan Liu, Hong Liu, and Chen Chen. Clickdiff: Click to induce semantic contact map for controllable grasp generation with diffusion models. 2024. 3

[48] Yong-Lu Li, Liang Xu, Xijie Huang, Xinpeng Liu, Ze Ma, Mingyang Chen, Shiyi Wang, Haoshu Fang, and Cewu Lu. Hake: Human activity knowledge engine. *ArXiv*, abs/1904.06539, 2019. 2, 3, 5

[49] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3, 6

[50] Xueyi Liu and Li Yi. Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. *ArXiv*, abs/2402.14810, 2024. 3

[51] Yunze Liu, Yun Liu, Chen Jiang, Zhoujie Fu, Kangbo Lyu, Weikang Wan, Hao Shen, Bo-Hua Liang, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20981–20990, 2022. 3

[52] Xiaoxiao Long, Yuanchen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8

[53] Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao. Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. *arXiv preprint arXiv:2311.17957*, 2023. 2, 6, 7, 8

[54] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Grasping diverse objects with simulated humanoids, 2024. 3

[55] Chaerin Min and Srinath Sridhar. Genheld: Generating and editing handheld objects. *arXiv preprint arXiv:2406.05059*, 2024. 3

[56] Chong Mou, Xintao Wang, Liangbin Xie, Jing Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *ArXiv*, abs/2302.08453, 2023. 2, 4, 13

[57] Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, and Minh Hoai. Handiffuser: Text-to-image generation with realistic hand appearances. In *CVPR*, pages 2468–2479, 2024. 2, 3

[58] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 13

[59] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 1

[60] 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, 4, 7, 14

[61] Anton Pelykh, Ozge Mercanoglu, and Richard Bowden. Giving a hand to diffusion models: A two-stage approach to improving conditional human image generation. *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10, 2024. 2

[62] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[63] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4331–4340, 2019. 4

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 4

[65] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 1

[66] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 1, 3, 4, 6, 7, 13

[67] 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), 2017. 2, 3, 4, 6

[68] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 7

[69] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022. 1

[70] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, Lille, France, 2015. PMLR. 1, 3

[71] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1

[72] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *The European Conference on Computer Vision (ECCV)*, 2020. 3, 4, 6, 7

[73] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13253–13263, 2021. 3, 4, 7, 13

[74] Purva Tendulkar, D'idac Sur'is, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21179–21189, 2022. 3, 4, 7, 13

[75] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 1, 8, 13

[76] Julen Urain, Niklas Funk, Jan Peters, and Georgia Chalvatzaki. Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffu-

sion. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5923–5930, 2022. 3

[77] Chengrui Wang, Pengfei Liu, Min Zhou, Ming Zeng, Xubin Li, Tiezheng Ge, and Bo zheng. Rhands: Refining malformed hands for generated images with decoupled structure and style guidance. *arXiv preprint arXiv:2404.13984*, 2024. 2

[78] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11359–11366, 2022. 3

[79] Zehang Weng, Haofei Lu, Danica Kragic, and Jens Lundell. Dexdiffuser: Generating dexterous grasps with diffusion models. *ArXiv*, abs/2402.02989, 2024. 3

[80] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image, 2024. 8

[81] Y. Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision*, 2021. 3, 13

[82] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 13

[83] Zihui Xue, Mi Luo, Chen Changan, and Kristen Grauman. Hoi-swap: Swapping objects in videos with hand-object interaction awareness. *arXiv preprint arXiv:2406.07754*, 2024. 3

[84] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4212–4222, 2023. 2

[85] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 8

[86] Yufei Ye, Xueting Li, Abhi Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22479–22489, 2023. 2, 3, 6, 7

[87] Zheng-Lun Yu, Shaoli Huang, Chengjie Fang, T. Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12955–12964, 2023. 6

[88] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2, 6, 7, 8

[89] Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, and Xiaolong Wang. Hoidiffusion: Generating realistic 3d hand-object interaction data. In *CVPR*, pages 8521–8531, 2024. 2, 3

[90] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 585–594, 2023. 3

[91] Yanzhao Zheng, Yunzhou Shi, Yuhao Cui, Zhongzhou Zhao, Zhiling Luo, and Wei Zhou. Coop: Decoupling and coupling of whole-body grasping pose generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2163–2173, 2023. 3, 4, 7, 13

[92] 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*, 2022. 3

[93] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*, 2024. 6, 7

[94] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 6

Supplementary Material for GraspDiffusion: Synthesizing Realistic Whole-body Hand-Object Interaction

S1. Model Architecture

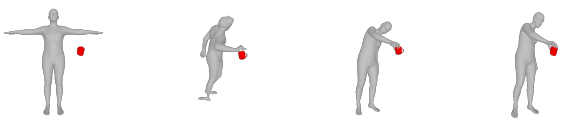
For the first stage of our pipeline, we use a diffusion model to synthesize a body pose grasping the input object. The model is trained to predict plausible body parameters (6 DoF body pose, global orientation), conditioned on the object's relative location $t_{\text{obj}} \in \mathbb{R}^3$ and the target hand $c_{\text{left}}, c_{\text{right}} \in \{0, 1\}$. These conditions are transformed to a conditional embedding v_c, which is added to the timestep embedding e_t and passed to residual blocks within the model, following [58]. We used 3 ResNet blocks for the model, and adopted a cosine noise schedule in training. Additional grasping results are provided in Figure 13, and details on model parameters are provided in Table 5.

In Figure 10, we provide an example comparison between our method and previous grasping pose generation methods [73, 74, 81, 91]. When given a object with its location relative to the human body (left row), GOAL [73] and SAGA [81] tend to create distorted poses when the object is far away from the human, as they assume it to be in the same horizontal xy-plane. FLEX assumes a world-centric coordinate system, which leads to pose ambiguity for our scenario. While COOP has a similar objective, it focuses on various object heights, and requires an extensive test-time optimization of 5 different loss terms. We are the first to utilize a lightweight diffusion model in synthesizing body grasping poses.

For the second stage of our pipeline, we use a diffusion model based on the Latent Diffusion [66] architecture, and attach encoders [56] that receives spatial features from the synthesized body pose. Specifically, we first provide three spatial conditions from the full-body grasping pose; the human skeleton projection, joint depth map, and the occluded object with ambient lighting ($[s^i, d^i, o^i]$). Then we further refine the hand-object region by providing similar spatial conditions, but centered on the hand region ($[s^i_h, d^i_h, o^i_h]$).

Parameter	Diffusion Model (Body Pose)
Input Channels	132
Condition Channels	5
Model Channels	1024
ResBlock Number	3
Diffusion Steps	1000
Noise Scheduler	Cosine

Table 5. Model architecture for body pose generation diffusion model.



(a) Location (b) GOAL [73] (c) COOP [91] (d) Ours

Figure 10. Grasp synthesis comparison with previous methods.

Figure 11. Synthesis results from a single image. We first synthesize a 3D Mesh from the image using TripoSR [75], InstantMesh [82], Real3D [31] and subsequently used the mesh as input.

Parameter	Conditional Encoder(s)
Input Channels	3×64
Output Channels	$[320, 640, 1280, 1280]$
ResBlock Number	2
Kernel Size	1
Feature Weight (Body)	$[1.0, 0.6, 1.0]$
Feature Weight (Hand)	$[1.0, 0.6, 1.0]$
Parameter	**Attention Injection**
Human Strength	0.2
Object Strength	1.8
Negative Object Strength	-9.0
Weight coefficient (w')	0.4

Table 6. Model architecture for scene generation models, and inference parameters for attention injection.

Figure 12. Failure cases for GraspDiffusion.

For training, we only train the conditional encoders and fix the parameters for the original Stable Diffusion model, encouraging the encoders to be used with other diffusion models finetuned from Stable Diffusion, accounting to our pipeline's style flexibility.

During inference, we further control the interaction by rectifying the cross-attention maps for the human and object. For the segmentation masks from the body pose (human : m^i, object : m^i_o, negative object : m^i_{no}), we assign different levels of strength for each maps to create an input attention matrix $A \in \mathbf{R}^{N_i \times N_t}$, where N_i and N_t are the number of image and text tokens. We assign a higher weight for the object masks due to their regional size differences. We then edit the cross attention layers so that it computes the output as $\text{softmax}(\frac{QK^T + wA}{\sqrt{d_k}})V$, where Q, K, V are the query, key and value embeddings, d_k is the dimensionality of Q and K, and w is a scalar weight that controls the total strength of user input attention. This encourages the image tokens in the segmented regions to adhere more to the corresponding text tokens, ensuring that the interaction captured by the body pose is well maintained. Following [4], we calculate w as

$$w = w^{'} \cdot \log(1 + \sigma) \cdot \max(QK^T)$$

where $w^{'}$ is a user defined scalar. Details on model parameters and inference are provided in Table 6. Note that for the feature weights, we assigned a relatively low rate for the joint depth map, to ensure the result image doesn't overfits to the SMPLX [60] mesh's outline.

S2. Additional Results

We display failure cases for our pipeline in Figure. 12, where We note some failure cases, where the refined hand stands out from the image (left row), the hand shape tends to be uncanny (middle row), or where the complex object texture is not correctly preserved within the image (right row).

We also provide additional results for realistic, full-bodied human object interaction image generation in Figure. 13. We display that our model is capable of producing images of realistic humans interacting with the given object, with high diversity over human identity, body pose, camera angle, background, and other relevant scene context. The results demonstrate our pipeline's capability in creating realistic grasps for unseen objects.

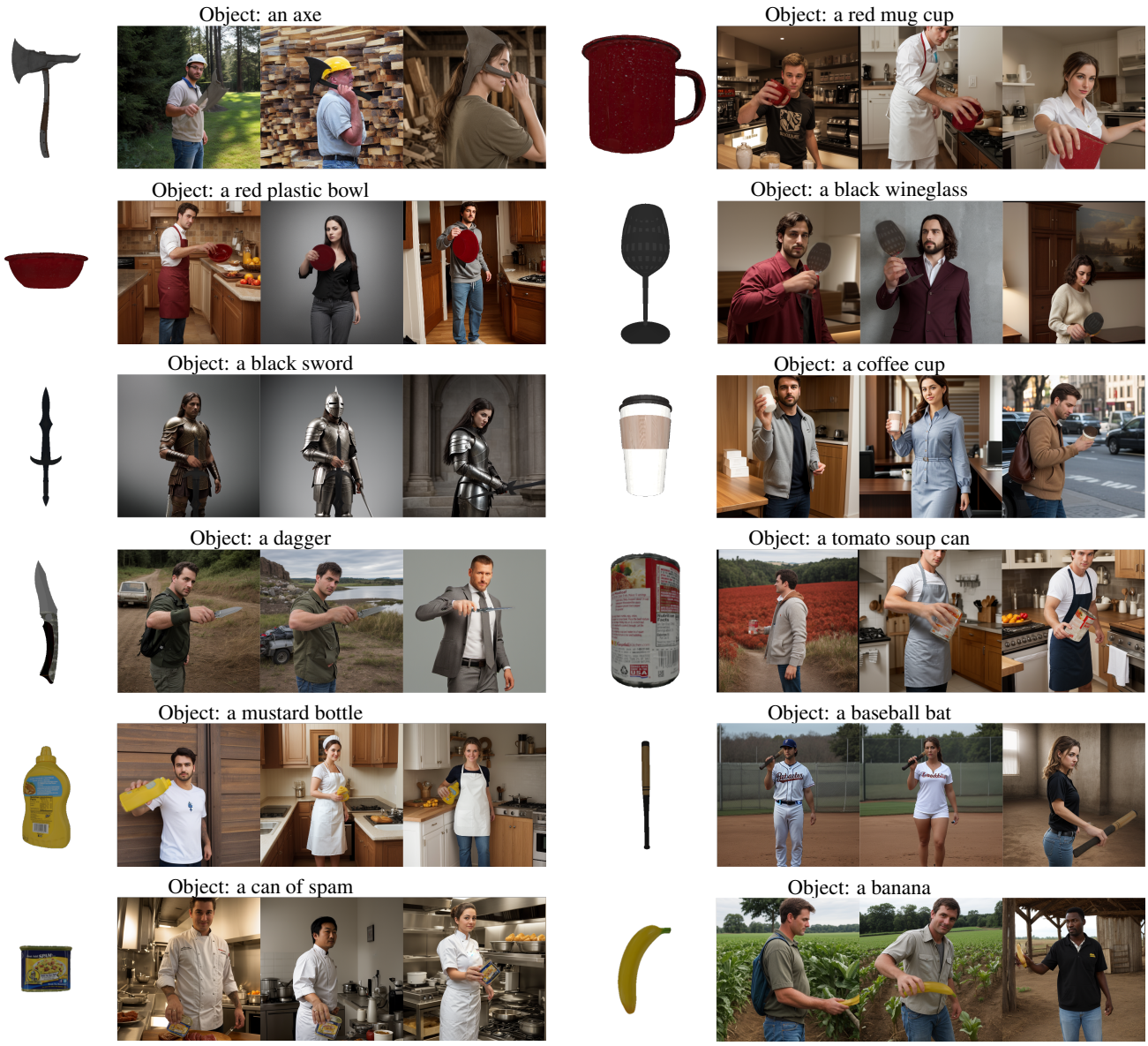


Figure 13. More results on synthesized images from a given object. Objects were gathered from the DexYCB dataset [12] and SketchFab [3]