HOIDiffusion: Generating Realistic 3D Hand-Object Interaction Data

Mengqi Zhang[1*] Yang Fu[1*] Zheng Ding[1] Sifei Liu[2]
Zhuowen Tu[1] Xiaolong Wang[1]
[1]UC San Diego [2] NVIDIA

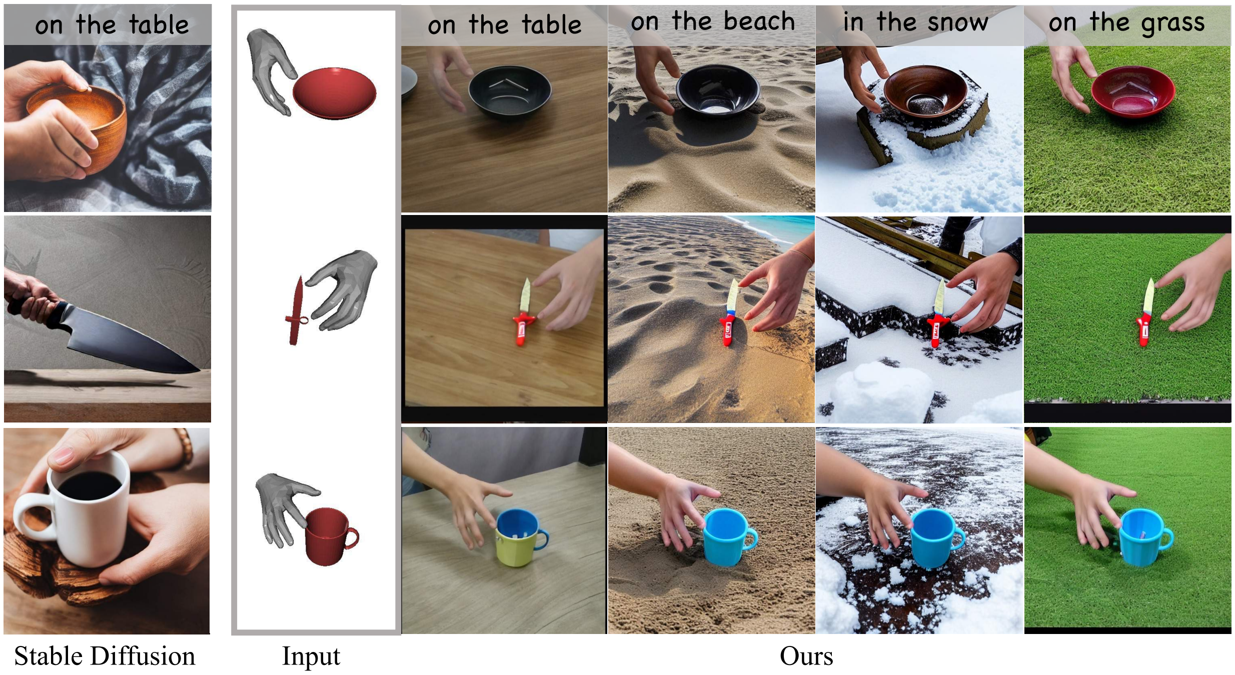


Figure 1. (i) **Left**: Hand-object synthesis with Stable Diffusion model; (ii) **Right**: HOIDiffusion generates high-quality hand-object interaction images conditioned on physical structures and detailed text description. The model disentangles the geometry from appearance, exhibiting high generation diversity. Each row: We can fix the structure and control the style based on text inputs; Each column: We can fix the style and control the structure based on 3D structural inputs.

Abstract

3D hand-object interaction data is scarce due to the hardware constraints in scaling up the data collection process. In this paper, we propose HOIDiffusion for generating realistic and diverse 3D hand-object interaction data. Our model is a conditional diffusion model that takes both the 3D hand-object geometric structure and text description as inputs for image synthesis. This offers a more controllable and realistic synthesis as we can specify the structure and style inputs in a disentangled manner. HOIDiffusion is trained by leveraging a diffusion model pre-trained on large-scale natural images and a few 3D human demonstrations. Beyond controllable image synthesis, we adopt the generated 3D data for learning 6D object pose estimation and show its effectiveness in improving perception systems. Project page: https://mq-zhang1.github.io/HOIDiffusion.

1. Introduction

Understanding how human hands interact with objects has been a long-standing problem in computer vision. Recently, researchers have tried to scale up such understandings by collecting videos on hand-object interactions [10, 17]. Models trained with these datasets focus on performing hand-object relational reasoning in 2D space. To enable broader applications in robotics and VR/AR, more efforts have been spent on collecting hand-object interaction data with 3D annotations via multiple cameras [6] and new labeling approaches with prepared object CAD models [19]. However, such a data collection process is not scalable and most datasets only contain dozens of objects.

Given the recent advancement of generative modeling with diffusion process [25, 53], can we leverage them to generate realistic 3D hand-object interaction data? While

[*] Equal Contribution.

state-of-the-art diffusion models such as Dall-E2 [44] and Stable Diffusion [46] can generate realistic images given text instructions, they still fail quite often when it comes to capturing the details of how fingers are placed around the object. As shown on the left side of Figure 1, Stable Diffusion might not be able to output physically or geometrically plausible interactions and sometimes there are more than five fingers in a hand. Moreover, it is still unclear how to configure image outputs beyond text instructions and make the output correspond to 3D shapes and poses.

In this paper, we propose to generate 3D hand-object interaction data, i.e., realistic images come with 3D groundtruths at the same time. Beyond realistic generation, we also enable controllable synthesis where the users can specify the geometry configuration and appearance in a disentangled manner. We achieve this by introducing a two-stage framework: We first synthesize the 3D geometric structure (shape and pose) of the hand and the object, and then we train a diffusion model conditioned on both the 3D structure and the text (indicating the style) to synthesize the corresponding RGB image. We visualize some synthesis results on the right side of Figure 1. In each row we generate images with the same 3D structural configuration but with different object and background styles; In each column, we fix the styles and synthesize images with different geometric structures.

In the first stage of our framework, we generate a human grasp based on a given 3D object model. We apply the pre-trained GrabNet [55] for this task, which takes the object mesh as inputs for a Variational AutoEncoder and predicts different grasp poses as outputs. In the second stage, we train a diffusion model conditioning the hand-object geometric configurations. We fine-tune the pre-trained Stable Diffusion model [46] with a few human demonstrations from the DexYCB dataset [6]. We convert the hand-object geometry to estimated surface normals, segmentation, and hand keypoint 2D projection as conditional inputs for the new diffusion model, specifying the structure of the image to generate. The diffusion model will also take text inputs for specifying appearance. During training, we apply a background regularization strategy to reduce the bias brought by DexYCB which comes with the same clean background. The fine-tuned diffusion model leverages both the rich appearance information from the pre-trained model and the geometry information from the new conditional variables.

In our experiments, our method outperforms previous approaches on hand-object image synthesis with more physically plausible interactions. The disentangled design provides flexible control of geometry and appearance. Interestingly, the model shows a strong generalization ability to different text prompts when changing the foreground and background appearance. We use several metrics to evaluate generation fidelity to real datasets and visual alignment with provided prompts. The results show an improvement compared to baselines. Additionally, with a generated dataset with both images and corresponding 3D geometry using our pipeline, we can use it to train an object pose estimator as a downstream application. Our experiments indeed show such realistic synthesized data is very helpful in improving the perception metrics.

2. Related work

Hand-Object Interaction Dataset. The understanding of hand-object interactions has been a long-standing problem [1, 18, 39, 40, 54]. More recently, the data-driven approaches [5, 12, 20, 21, 31, 38] have shown significant advancement in hand-object shape reconstruction and pose estimation. At the heart of this progress, is the collection of hand-object interaction data. To obtain 3D annotations, existing datasets [13, 19, 20, 55, 61] collect videos with attached sensors or mocap markers to track hand pose or utilize optimized algorithms to facilitate annotations. 2D annotations are also provided with manual labeling [6]. However, these approaches are time-consuming not scalable. Recently more data collection pipelines [16, 32, 42, 60] have taken advantage of detection and segmentation techniques to automatically acquire annotations. However, even though this method eases the difficulty, estimations may not be precise enough, and manual annotations might still be required for accurate 3D annotations. The diversity of the data is also relatively small given the repetitive patterns in videos. In this paper, we propose a new effective data generation method facilitated by generative models for hand-object interaction images with full 3D annotations.

Hand Grasp Generation. Hand grasp generation given an object model [2, 8, 9, 16, 26, 27, 63, 65] is of vital importance in our method to provide an ending pose for grasping trajectory. Most approaches estimate or further refine the grasping hand pose by predicting the contact map between hands and objects [3, 16, 26]. Other methods [13, 27, 55] predict the MANO parameter hand representation introduced in [47] with variational autoencoder or implicit function architectures. Additionally, The hand parameter prediction can be integrated as a component of the whole human body [56]. By taking advantage of these grasp predictors, we are able to obtain satisfying ending poses for hand trajectory generation.

Diffusion Models. Diffusion models [25, 46, 53] which learn to denoise images from Gaussian distributions, emerge recently and perform photo-realistic image synthesis with more stable training process compared to other generative models [15, 28]. Many successive advancements occur in this field [11, 36, 37, 44, 50]. More relevantly, special tokens are introduced [14, 49] to fine-tune the model, enabling personalized text-to-image generation. With these

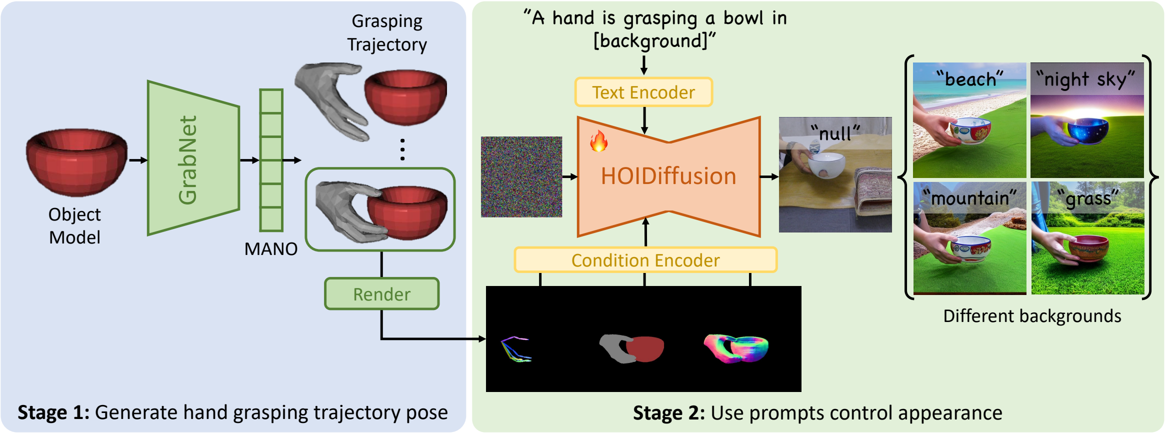


Figure 2. **Pipeline.** We propose a two-stage pipeline to synthesize hand-object-interaction data. During the first stage, we utilize a pretrained GrabNet to output 3D hand poses given by a single object model. Then in the second stage, we use those 3D hand poses along with segmentation maps, normal maps and skeletons to conditionally generate high-quality HOI data.

amazing generation results, researchers [4, 59] start investigating the possibility of utilizing diffusion models as a new source of data for classification or segmentation tasks. Besides using text inputs, conditional generation with a given layout or scratch achieves impressive performance after the appearance of CoAdapter [33] and ControlNet [64]. These works unveiled the large potential of diffusion models to control or edit images according to users' demands. However, these models still suffer from generating realistic hand-object-interaction images, such as not being object-agnostic, inaccurate geometry, and generating hands with missing fingers or unnatural poses. In our work, we focus on how to utilize physical conditions, including normal, hand skeleton projection and segmentation to construct a 3D-aware model for scalable hand-object-image generation with flexible geometry and appearance control.

3. Method

In order to generate scalable hand-object-interaction images, three expectations need to be satisfied: (i) The model should generate realistic images that are consistent with the geometric description of the specified object. (ii) It should retain the prompt-editing capabilities inherent in the stable diffusion model while incorporating controllable conditions. (iii) The model should have a better generalization ability to synthesize images of unseen instances or categories.

To meet the above requirements, we propose a two-stage approach. For the first stage, our goal is to establish the conditions, primarily the hand-grasping trajectory, for the subsequent stage. Specifically, we utilized a generalizable VAE model trained on large-scale 3D physical data to obtain the ending pose and interpolated the trajectory using spherical linear interpolation. For the second stage, we extracted multiple geometry structures either from the generated grasping trajectory or from images with natural hand-grasping actions through rendering and off-the-shelf estimators to finetune a controllable Stable Diffusion, which enables precise pose control at inference. Furthermore, a background regularization strategy is introduced in our pipeline to mitigate edit ability degradation brought by finetuning. The entire pipeline is shown in Figure 2.

3.1. Preliminary

Denoising diffusion model [25] is a kind of new generative model with competitive performance and more stable training. It consists of two main processes: diffusion and denoising. In the forward process, randomly sampled noises are added to original images, which can be mathematically simplified as $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)I)$, where α_t is hyperparameters control noise scheduling, and $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. After T steps, the distribution shifted from image space to approximate standard Gaussian distribution. And U-Net model is trained to predict the added noise. During inference, an image is initialized from the Gaussian distribution and the model removes the noise within T steps, with each step t formulated as

$$\hat{\epsilon}_t = f_\theta(x_t, t) \tag{1}$$

where f_θ is the U-Net model. The estimated image at the next time step can be derived and written as

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\hat{\epsilon}_t) + \sigma_t z \tag{2}$$

The simpler version of training loss is:$||\epsilon - f_\theta(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)||^2$. Therefore diffusion models are able to generate photorealistic images.

3.2. Hand Grasping Trajectory Generation

In the first stage of our framework, given a randomly transformed object model, we require a hand-grasping trajectory to effectively reach the object. To this end, we adopted an interpolation method using GrabNet, a model for generating hand grasps conditioned on BPS (Body Part Segmentation) [41] derived from object models. Through extensive training on a large dataset, it has learned to accurately map contact between hands and objects, showcasing strong generalization capabilities for unseen scenarios. Utilizing GrabNet, we can determine the final grasp hand MANO parameters, which include joint pose and hand shape. Given the ending pose and the object's position, we can approximate an initial starting point—positioned vertically above the contact surface and within a certain distance—using zeroed MANO parameters. Subsequently, we employ spherical linear interpolation between this starting point and the ending pose to create a smooth hand-grasping trajectory. As shown in Figure 2 Stage 1, the grasping parameters can be obtained along the trajectory and we could acquire different ground-truth annotations such as segmentation, mask, and depth maps through rendering.

3.3. Hand-Object-Interaction image Synthesis

Our Hand-Object-Interaction image synthesis module mainly consists of two components: structure control and appearance regularization, to disentangle the geometry and appearance separately.

Structure Control Given the hand-grasp trajectory, we extract multiple geometric conditions and leverage the advanced Stable Diffusion models to generate realistic images that are consistent with the given conditions. To precisely control the hand-object image generation, three structure conditions are distilled to control the generation. i) To synthesize realistic hands without missing fingers which is the problem encountered in original Stable Diffusion, hand position information is essentially required. Instead of directly using MANO parameter vectors as guidance, we projected the skeleton information onto the image space as visual control, which can be denoted as s^i. ii) Additionally, to mitigate the inter-disturbance of hand object areas and degradation in performance brought by occlusion, hand-object segmentation (m^i) is used to provide clear boundaries to separate areas, and coarse object shape priors. iii) Finally, we also apply an estimated normal map (n^i) to seize the surface geometry with lighting. The forward process defined in Equation 1 can now be defined as $\hat{\epsilon_t^i} = f_\theta(x_t^i, t, [s^i, m^i, n^i])$. With the above controls, the structure information could be disentangled from the appearance, and thus during the inference, we could seamlessly synthesize an accurate image with new poses.

Appearance Regularization With the above component, we are capable of synthesizing images aligned with diverse condition geometries during inference. However, a notorious drawback of fine-tuning is its tendency to converge or overfit quickly to the training dataset's style, thereby significantly reducing image diversity. To mitigate this problem and fully harness the capabilities of text-to-image diffusion models for flexible style transformation via prompts, we introduce an appearance regularization method combined with classifier-free guidance [24] as shown in Figure 3. Specifically, in addition to using the original Hand-Object Interaction (HOI) training dataset, we synthesize batches of high-quality scenery images with the pre-trained text-to-image diffusion model. The prompts for these images are generated by the large language model ChatGPT, forming what we refer to as a "background buffer". During training, we intermittently utilize these background images for regularization, ensuring it does not detrimentally impact performance. Meanwhile, the paired blank conditions are applied as classifier-free guidance, corresponding to the background region in HOI data. The objective becomes:

$$\begin{aligned} \mathcal{L} = E_{x_0, x_r, \epsilon, \epsilon_r}[||\epsilon - f_\theta(\sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)||^2 \\ + w_r||\epsilon_r - f_\theta(\sqrt{\overline{\alpha}_t} x_r + \sqrt{1-\overline{\alpha}_t}\epsilon_r, t)||^2] \end{aligned} \tag{3}$$

where w_r is the regularization weight, which is set to 1 in our experiments. x_r, ϵ_r are input from the background buffer and corresponding added noise. In addition to the buffer, we also use the large multimodal model LLaVA [30] to caption our training images with detailed descriptions of foreground appearance and background, forcing the model aware of diverse texts with the assistance of the CLIP [43] text encoder.

4. Experiment

4.1. Implementation Details

To achieve scalable synthesis and improve generalization ability, the training dataset is required to encompass diverse backgrounds and comprehensive annotations. For these purposes, the DexYCB dataset is selected for its high-quality images from varied viewpoints of hand gestures, with human demonstrations and diverse background settings. To prevent overfitting and preserve the text-driven editing ability of diffusion models, we adopt the learning rate of 10^{-5} during training. About 50,000 steps prove sufficient to achieve satisfactory generation quality. When utilizing images from background buffers, we provide vacant skeleton projection and mask conditions, corresponding to the background regions in HOI training data images. This design also can be viewed as a classifier-free guidance that drives the model aware of the condition control effects. The entire training process costs approximately 12 hours on eight A100 GPUs.

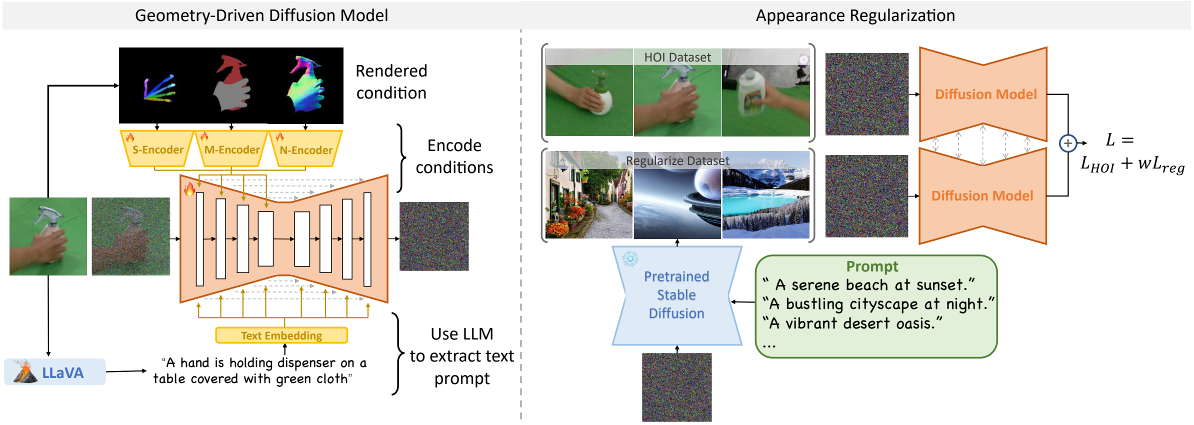

Geometry-Driven Diffusion Model
Appearance Regularization
Rendered condition
S-Encoder
M-Encoder
N-Encoder
Encode conditions
Text Embedding
LLaVA
"A hand is holding dispenser on a table covered with green cloth"
Use LLM to extract text prompt
HOI Dataset
Regularize Dataset
Diffusion Model
Diffusion Model
$L = L_{HOI} + wL_{reg}$
Pretrained Stable Diffusion
Prompt
" A serene beach at sunset."
"A bustling cityscape at night."
"A vibrant desert oasis."
...

Figure 3. **Model Figure.** We inject three conditional encoders into the stable diffusion model. We utilize both the HOI datasets and high-quality background images to train HOIDiffusion. The background images are synthesized using the scenery prompts. The texts sent to the model are output by LLaVA for detailed description.

4.2. Evaluation and Baseline Comparison

Baselines We compare the results with four baseline models customized for our setting: (1) a fine-tuned LDM [46] without condition modules, working as an unconditional generation task; (2) a modified DreamBooth [49] fine-tuned model, with proposed specific token agnostically applied to all object categories, complemented by regularization data from our trained conditional stable diffusion; (3) Affordance Diffusion [62]; (4) ControlNet with same multiple condition input. It is important to mention that the Affordance Diffusion model is only used for comparison on contact recall evaluation, with quantitative results reported from the original paper.

Image Synthesis Quality To assess the generation quality, we adopt commonly used metrics FID [23] to evaluate the fidelity of our generated images to real datasets. Additionally, we also present the comparison results with baselines on the Inception Score [51] and sFID [35]. Our reference batch consists of 50k randomly selected images from the training dataset in total. For the sample batch, we synthesize hand-grasping images corresponding to all randomly rotated object models, some of which are unseen during training. The total sample size is 5k. We refer readers to Table 1 for a comprehensive overview of the comparison.

Appearance Alignment Furthermore, to demonstrate the flexible control over object and background appearance, we use another metrics CLIPScore [22] originally designed for image captioning to automatically evaluate the alignment level between generated images and corresponding prompts. We randomly sample 50 instances from all object models with fixed structure conditions. Multiple appearance descriptions generated by ChatGPT are applied to each instance. The results are shown in Table 1. Notably, our method demonstrates superior overall quality of generated images, with higher fidelity to real data and improved alignment with appearance-controlling texts.

Method	FID↓	sFID↓	IS↑	CLIPScore ↑
LDM [46]	63.71	119.10	6.81	0.68
DreamBooth [49]	134.40	92.82	**7.99**	0.75
ControlNet [64]	87.99	248.56	6.60	0.77
HOIDiffusion (Ours)	**55.22**	**91.28**	7.73	**0.78**

Table 1. **Quantitative comparison with previous baseline methods**. All models are trained on the DexYCB. We use FID to directly measure the synthesis quality of generated hand-object interaction images. sFID is a recently proposed metric to evaluate image quality using higher-level spatial features. IS is measured for diversity and CLIPScore is to evaluate generated images alignment with provided prompts.

Hand Pose Evaluation For hand-object-interaction images, hand grasping status, and pose precision are of vital importance for real-world applications. A fundamental criterion for synthesized images is the geometric consistency between our generated hands and the provided 2D skeleton projection, disregarding the depth dimension. On top of that, these grasping images sometimes serve as visual demonstrations for various downstream tasks, requiring the exactly accurate contact status between the hand and object in the ending pose image along the grasping trajectory. Consequently, our model is evaluated with the baselines on two key perspectives: Hand contact recall and hand re-inference accuracy. Specifically, we adopt a contact evaluation setup utilized in Affordance Diffusion [62]. An off-the-shelf hand-object detector [52] is used to classify the image's in-contact status. Furthermore, to evaluate the re-inference accuracy, we estimate the MANO parameters of hands in images through a widely used single-view hand pose estimator [48], from which we derive the predicted hand joint positions. The percentage of correct keypoints (PCK) is used to measure the accuracy of predicted keypoints representing the hand poses in our data. We evalu-

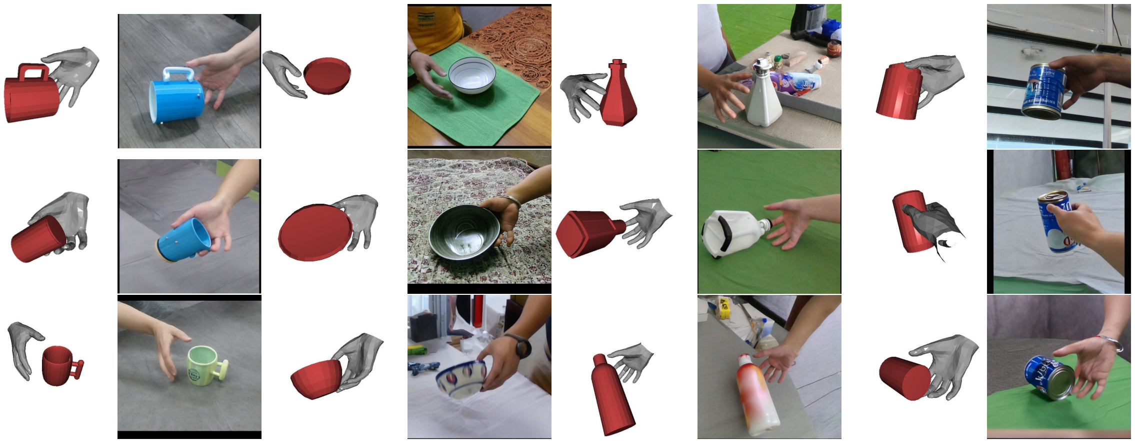


Figure 4. **Qualitative results on different structures.** Generated images with the same background description but different physical conditions (object shape, poses, and hand skeletons). With plain prompts, HOIDiffusion could generate more realistic images similar to the style in training datasets.

Method	Hand Contact Recall %					PCK↑
	Mug	Bowl	Bottle	Can	Mean↑	
LDM [46]	79.12	78.89	73.00	60.64	72.91	0.15
DreamBooth [49]	79.12	62.22	78.00	75.53	73.72	0.10
ControlNet [64]	86.00	91.40	89.00	93.00	89.85	0.67
Affordance Diffusion [62]	73.00	90.00	90.00	-	84.33	-
HOIDiffusion (Ours)	92.31	97.78	94.00	97.87	**95.49**	**0.85**

Table 2. **Evaluation metrics from hand perspective.** Hand Contact Recall is to evaluate whether the hand-ending pose is in close contact with the object. PCK is used to measure the accuracy of generated images' keypoints

ate the pose precision on a subset of daily objects. As delineated in Table 2, benefited from the geometry guidance, our method manifests the capability to synthesize images depicting accurate grasps and firm contacts, outperforming previous methods in achieving higher mean contact recall and PCK.

4.3. Geometry and Appearance Disentangle

The most important design in our model is to disentangle the physical geometries from textures. Through this design, we have observed the remarkable ability of our model to control the generation process with novel object shapes and previously unseen text descriptions. During training, our model learns extensive shape priors from masks and normal map conditions, hence acquiring the capability to transfer to unseen object instances seamlessly. Furthermore, our model preserves the robust text editing ability, enabling flexible style transformation over both background and object appearance. In this subsection, we primarily focus on and showcase the qualitative results of structure and style manipulation.

Geometry Manipulation In Figure 4, we present the generated paired images of four daily seen objects: mug, can, bottle, and bowl, given varying instance shapes, poses, and hand skeletons. The left column of each pair is rendered image in 3D space, from which physical conditions are extracted. Normal maps are obtained from an estimated depth provided by the depth estimator MiDaS [45]. The skeleton and segmentation are also concurrently obtained during rendering. Corresponding generated images are displayed in the right column. All provided prompts follow the format: "A hand is grasping a [object]". The results exhibit an overall generation style in a laboratory or stereo environment, consistent with the realistic appearance style in the training dataset. From Figure 1, it is also evident that HOI images generated from our model are more closely aligned with the required geometry than baseline methods.

Background and object appearance control In this section, we explore the text editing ability with fixed geometry. Through background regularization and classifier-free guidance, our model exhibits the ability to depict diverse background contents, retaining control over appearance using text prompts. We investigate the qualitative performance of various text controls under identical physical conditions, shown in Figure 5. Each column represents one distinct background description. Notably, the generated images exhibit high fidelity to the provided prompts, maintaining the layout and structure unchanged. This demonstrates the ability of HOIDiffusion to effectively disentangle the appearance from geometry structures, thus enabling flexible style transformation without geometry distortion. This is essential in data construction, ensuring the precise alignment with input geometry and diverse range of visual appearances.

4.4. Applications

Video Generation The real hand-object interaction datasets often exhibit data in video format, a complete fetching process to the object. Collecting these video clips is a considerable challenge. Some datasets [6, 60] comprising almost

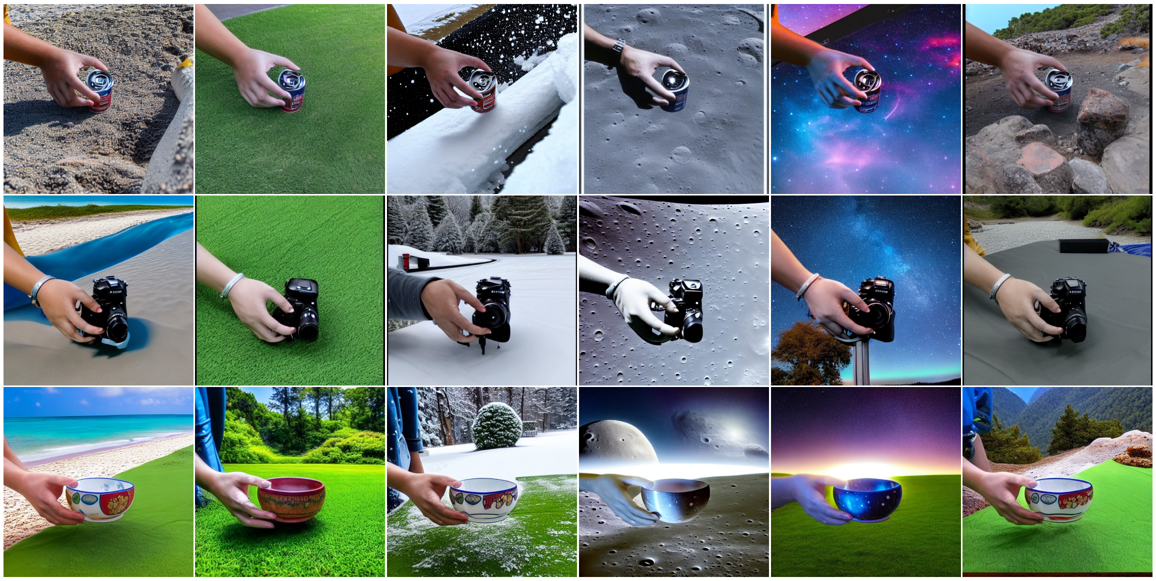


Figure 5. **Synthesized images with diverse background descriptions.** In addition to real-style synthesis, our model also allows users to generate according to their preferences such as science fiction or general landscapes.

millions of images only contain no more than 10k videos, thus video data is much more valuable. In experiments, we observe the significant divergence in generated images between adjacent frames despite the similar provided conditions and texts. This divergence results in dramatic flickering in directly concatenated videos. To this end, we leverage zero-shot video generation techniques in the diffusion model to establish inter-consistency among frames. To be more specific, the original self-attention layers in the U-Net are refactored to cross-attention modules between an anchor frame and current frames, This adjustment establishes the awareness of the previous appearance style, ensuring video consistency. In our experiments, we set the middle frame as an anchor, and with all frames attend to both the anchor image and themselves. This approach effectively mitigates the flickering issue, synthesizing relatively smooth hand-grasping trajectories. The video clip samples are shown in Figure 6.

Downstream tasks Another interesting method to evaluate the performance of HOIDiffusion, is to apply it in downstream tasks as a data augmentation method or new data source. In this section, we explore the potential for improving categorical object 6D pose estimation tasks. Most models in this task are trained on dataset NOCS [58], consisting of both synthetic and real data with annotations. Some models [7, 57] implicitly predict normalized object coordinate space, and then utilize Umeyama algorithms to parse the transformation matrix. Others [29] explicitly predict the rotation, translation, and scale parameters. Despite differing in module design, all these methods leverage RGB image encoders to obtain the visual features. An interesting observation is that the synthesized images directly rendered from object models appear too artificial, potentially affecting the performance of the RGB encoder. Inspired by this, we substitute the synthesized images with our generated HOI images in the same poses, anticipating the reality brought by our data could help enhance model performance. We exhibit the results in Table 3. We choose two representative object pose estimators for evaluation. The "original" model in the table refers to training using an unchanged NOCS dataset, and "our" model is trained using our mixed data.

5. Ablation Study

Two indispensable components of our design are precise structural control and appearance regularization, effectively improving model performance on geometry consistency and diversity.

Structural Control In our approach, there are three crucial structure conditions provided to the model and we investigate the importance brought separately by these three modules. Results are presented in Table 4. Essentially, each condition serves a distinct purpose: the normal map guides the model in perceiving surface textures with lighting, which is essential for maintaining geometry consistency; hand keypoint projection precisely depicts the pose of hand joints, preventing the model from synthesizing multiple fingers or distorted hands; hand-object segmentation provides a clear boundary between different regions, avoiding interference between hand and object areas. As presented in the results, our full model outperforms other incomplete versions in quantitative evaluation.

Appearance Regularization Table 5 and Figure 7 demonstrate the significance of appearance regularization in our module. Without regularization, the finetuned model

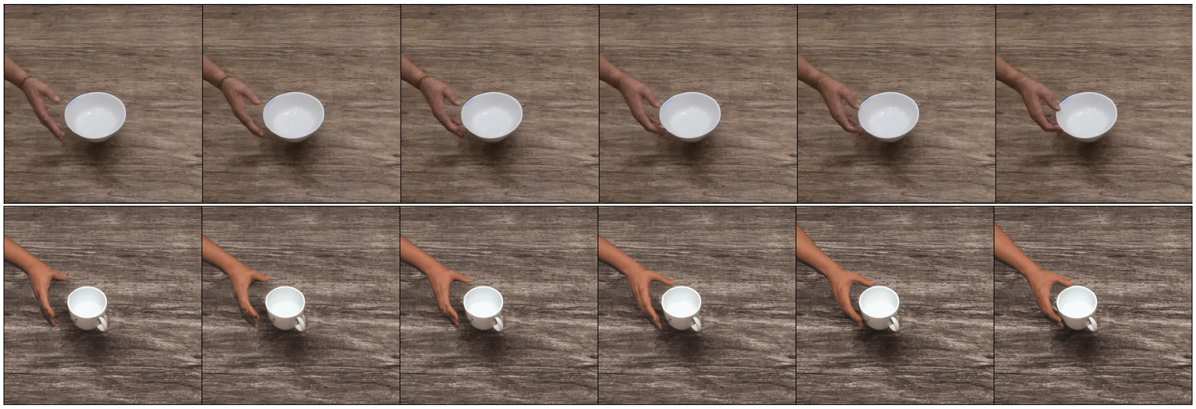


Figure 6. **Zero-shot video generation of hand grasping trajectory.** Images along the same line represent the sequential motion of reaching an object. By leveraging temporal-level cross-attention, the frame flickering problem is mitigated.

Method		IoU@25	IoU@50	IoU@75	5°2cm	5°5cm	10°2cm	10°5cm
SPD	Original	**82.9**	**75.3**	**48.6**	17.7	19.9	38.8	48.3
[34]	Ours	82.5	71.1	47.0	**21.1**	**23.2**	**43.8**	**54.5**
DualPoseNet	Original	84.1	79.7	60.1	28.0	34.3	47.8	64.2
[29]	Ours	**90.9**	**84.1**	**65.8**	**29.2**	**34.4**	**55.0**	**66.6**

Table 3. **Quantitative evaluation on NOCS.** We use SPD and DualPoseNet and change the synthesized images in the dataset with our generated images for training. Our performance improve on all metrics with DualPoseNet and all cm metrics with SPD which demonstrates the good quality of our images and can be utilized for downstream tasks.

Method	FID (1k) ↓
LDM (finetuned)	86.12
w/o estimated normal maps	82.40
w/o hand keypoint projection	81.93
w/o hand-object segmentation	78.57
HOIDiffusion(Ours)	**77.64**

Table 4. **Ablation study on structural control.** FID evaluation on 1,000 images of different types of missing modules to demonstrate the necessity of all physical conditions. Our method outperforms all others.

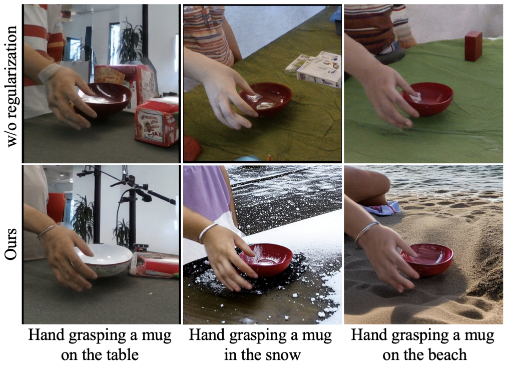
Method	CLIPScore ↑
w/o regularization	0.66
HOIDiffusion(Ours)	**0.79**

Table 5. **CLIPScore evaluation.** Consistency is evaluated between provided prompts and generated images for different backgrounds and instances.

quickly converges to the style in training datasets, mostly in a laboratory/studio environment, as depicted in the first line of Figure 7, This convergence impairs the model's ability to generate diverse images, which is essential for data generation. By incorporating appearance regularization using data generated from the pretrained model, HOIDiffusion mitigates the drift to a fixed style and improves the overall text-editing ability.

6. Conclusion

In this paper, we propose HOIDiffusion with precise appearance and structure control. We did experiments on geometry and appearance manipulation, and evaluated the performance using FID, IS, and hand contact recall. The results demonstrate better performance compared to baseline models. We apply the generated data for object 6D pose estimation and show its effectiveness in possibilities to improve perception systems.



Hand grasping a mug on the table

Hand grasping a mug in the snow

Hand grasping a mug on the beach

Figure 7. **Ablation study on appearance regularization**. Different backgrounds are used as prompts with the same geometry conditions to compare the text editing flexibility brought by the regularization module.

Acknowledgement Zhuowen Tu is supported by NSF IIS-2127544. Xiaolong Wang's lab is supported, in part, by NSF CAREER Award IIS-2240014, Qualcomm Innovation Fellowship, Amazon Research Award, and CISCO Faculty Award.

References

[1] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *European Conference on Computer Vision*, pages 640–653. Springer, 2012. 2

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

[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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 2

[4] Max F Burg, Florian Wenzel, Dominik Zietlow, Max Horn, Osama Makansi, Francesco Locatello, and Chris Russell. A data augmentation perspective on diffusion models and retrieval. *arXiv preprint arXiv:2304.10253*, 2023. 3

[5] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021. 2

[6] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 1, 2, 6

[7] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2773–2782, 2021. 7

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

[9] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 2

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 1

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[12] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6608–6617, 2020. 2

[13] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 2

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

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

[17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1

[18] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 671–678. IEEE, 2010. 2

[19] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 1, 2

[20] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 2

[21] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 571–580, 2020. 2

[22] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5

[23] 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 equilib-

rium. *Advances in neural information processing systems*, 30, 2017. 5

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 3

[26] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11107–11116, 2021. 2

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

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[29] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3560–3569, 2021. 7, 8

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 4

[31] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 2

[33] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 3

[34] Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-sdf: Learning disentangled signed distance functions for articulated shape representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13001–13011, 2021. 8

[35] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021. 5

[36] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

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

[38] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Generalized feedback loop for joint hand-object pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1898–1912, 2019. 2

[39] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *2011 International Conference on Computer Vision*, pages 2088–2095. IEEE, 2011. 2

[40] Paschalis Panteleris, Nikolaos Kyriazis, and Antonis A Argyros. 3d tracking of human hands in interaction with unknown objects. In *BMVC*, pages 123–1, 2015. 2

[41] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4332–4341, 2019. 4

[42] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022. 2

[43] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[45] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 6

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 5, 6

[47] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2

[48] 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. 5

[49] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 5, 6

[50] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[51] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 5

[52] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 5

[53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2

[54] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *European Conference on Computer Vision*, pages 294–310. Springer, 2016. 2

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

[56] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13263–13273, 2022. 2

[57] Meng Tian, Marcelo H Ang Jr, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 7

[58] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 7

[59] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[60] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20953–20962, 2022. 2, 6

[61] Ruolin Ye, Wenqiang Xu, Zhendong Xue, Tutian Tang, Yanfeng Wang, and Cewu Lu. H2o: A benchmark for visual human-human object handover analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15762–15771, 2021. 2

[62] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22479–22489, 2023. 5, 6

[63] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021. 2

[64] 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*, pages 3836–3847, 2023. 3, 5, 6

[65] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15741–15751, 2021. 2