Affordance Diffusion: Synthesizing Hand-Object Interactions

Yufei Ye\textsuperscript{1*} Xueting Li\textsuperscript{2} Abhinav Gupta\textsuperscript{1} Shalini De Mello\textsuperscript{2}
Stan Birchfield\textsuperscript{2} Jiaming Song\textsuperscript{2} Shubham Tulsiani\textsuperscript{1} Sifei Liu\textsuperscript{2}
\textsuperscript{1}Carnegie Mellon University \textsuperscript{2}NVIDIA

https://judyye.github.io/affordiffusion-www

1. Introduction

Consider the bottles, bowls and cups shown in the left column of Figure 1. How might a human hand interact with such objects? Not only is it easy to imagine, from a single image, the types of interactions that might occur (e.g., ‘grab/hold’), and the interaction locations that might happen (e.g., ‘handle/body’), but it is also quite natural to hallucinate—in vivid detail—several ways in which a hand might contact and use the objects. This ability to predict and hallucinate hand-object-interactions (HOI) is critical to functional understanding of a scene, as well as to visual imitation and manipulation.

Can current computer vision algorithms do the same? On the one hand, there has been a lot of progress in image generation, such as synthesizing realistic high-resolution images spanning a wide range of object categories \cite{43, 73} from human faces to ImageNet classes. Newer diffusion models such as Dall-E 2 \cite{65} and Stable Diffusion \cite{66} can generate remarkably novel images in diverse styles. In fact, highly-realistic HOI images can be synthesized from simple text inputs such as “a hand holding a cup” \cite{65, 66}.

On the other hand, however, such models fail when conditioned on an image of a particular object instance. Given

Recent successes in image synthesis are powered by large-scale diffusion models. However, most methods are currently limited to either text- or image-conditioned generation for synthesizing an entire image, texture transfer or inserting objects into a user-specified region. In contrast, in this work we focus on synthesizing complex interactions (i.e., an articulated hand) with a given object. Given an RGB image of an object, we aim to hallucinate plausible images of a human hand interacting with it. We propose a two-step generative approach: a LayoutNet that samples an articulation-agnostic hand-object-interaction layout, and a ContentNet that synthesizes images of a hand grasping the object given the predicted layout. Both are built on top of a large-scale pretrained diffusion model to make use of its latent representation. Compared to baselines, the proposed method is shown to generalize better to novel objects and perform surprisingly well on out-of-distribution in-the-wild scenes of portable-sized objects. The resulting system allows us to predict descriptive affordance information, such as hand articulation and approaching orientation.

\* Yufei was an intern at NVIDIA during the project.
an image of an object, it remains an extremely challenging problem to generate realistic human object interaction. Solving this problem requires (at least implicitly) an understanding of physical constraints such as collision and force stability, as well as modeling the semantics and functionality of objects — the underlying affordances [19]. For example, the hand should prefer to grab the kettle handle but avoid grabbing the knife blade. Furthermore, in order to produce visually plausible results, it also requires modeling occlusions between hands and objects, their scale, lighting, texture, etc.

In this work, we propose a method for interaction synthesis that addresses these issues using diffusion models. In contrast to a generic image-conditioned diffusion model, we build upon the classic idea of disentangling \textit{how} to interact (layout) from \textit{how to interact} (content) [25, 30]. Our key insight is that diverse interactions largely arise from hand-object layout, whereas hand articulations are driven by local object geometry. For example, a mug can be grasped by either its handle or body, but once the grasping location is determined, the placement of the fingers depends on the object’s local surface and the articulation will exhibit only subtle differences. We operationalize this idea by proposing a two-step stochastic procedure: 1) a LayoutNet that generates 2D spatial arrangements of hands and objects, and 2) a ContentNet that is conditioned on the query object image and the sampled HOI layout to synthesize the images of hand-object interactions. These two modules are both implemented as image-conditioned diffusion models.

We evaluate our method on HOI4D and EPIC-KITCHEN [11, 48]. Our model outperforms generic image generation baselines, and the extracted hand poses from our HOI synthesis are favored in user studies against baselines that are trained to directly predict hand poses. We also demonstrate surprisingly robust generalization ability across datasets, and we show that our model can quickly adapt to new hand-object-interactions with only a few examples. Lastly, we show that our proposed method enables editing and guided generation from partially specified layout parameters. This allows us to reuse heatmap prediction from prior work [13, 56] and to generate consistent hand sizes for different objects in one scene.

Our main contributions are summarized below: 1) we propose a two-step method to synthesize hand-object interactions from an object image, which allows affordance information extracted from it; 2) we use inpainting techniques to supervise the model with paired real-world HOI and object-only images and propose a novel data augmentation method to alleviate overfit to artifacts; and 3) we show that our approach generates realistic HOI images along with plausible 3D poses and generalizes surprisingly well on out-of-distribution scenes. 4) We also highlight several applications that would benefit from such a method.

2. Related Work

**Understanding Hand-Object-Interaction.** In order to understand hand-object-interaction, efforts have been made to locate the active objects and hands in contact in 2D space, via either bounding boxes detection [4, 53, 75] or segmentation [15, 76]. Many works reconstruct the underlying shape of hands and objects from RGB(D) images or videos by either template-based [6, 18, 26, 82, 84] or template-free methods [9, 28, 38, 90]. Furthermore, temporal understanding of HOI videos [20, 27, 62, 63, 81] aims to locate the key frames of state changes and time of contact. In our work, we use these techniques to extract frames of interests for data collection and to analyze the synthesis results. While these works recognize what is going on with the underlying hands and objects, our task is to hallucinate what hands could possibly do with a given object.

**Visual Affordance from Images.** Affordance is defined as functions that environments could offer [19]. Although the idea of functional understanding is core to visual understanding, it is not obvious what is the proper representation for object affordances. Some approaches directly map images to categories, like holdable, pushable, liftable, etc. [7, 30, 44, 57]. Some other approaches ground these action labels to images by predicting heatmaps that indicate interaction possibilities [13, 34, 47, 56, 59]. While heatmaps only specify \textit{where} to interact without telling \textit{what} to do, recent approaches predict richer properties such as contact distance [38], action trajectory [47, 54], grasping categories [22, 49], etc. Instead of predicting more sophisticated interaction states, we explore directly synthesizing HOI images for possible interactions because images demonstrate both \textit{where} and \textit{how} to interact comprehensively and in a straightforward manner.

**3D Affordance.** While 3D prediction from images is common for human-scene interaction [16, 23, 25, 45, 93, 94], 3D affordance for hand and object typically takes in known object 3D shapes and predicts stable grasps [24, 37, 52, 60]. In contrast, our work predicts \textit{coarse} 3D grasps from RGB images of \textit{general} objects. The coarse but generalizable 3D hand pose prediction is shown as useful human prior for dexterous manipulation [2, 12, 41, 49, 64, 88]. While some recent works [10, 38] generate 3D hand poses from images by direct regression, we instead first synthesize HOI images from which 3D hand poses are reconstructed afterwards.

**Diffusion Models and Image Editing.** Diffusion models [32, 79] have driven significant advances in various domains [42, 83, 89, 91, 95], including image synthesis [3, 65, 66, 72]. A key advantage of diffusion models over other families of generative models [21, 40] is their ability to easily adapt to image editing and re-synthesis tasks without much training [17, 39, 70]. While recent image-conditioned generative models achieve impressive results on various image translation tasks such as image editing [5, 35, 50],
3. Method

Given an image of an object, we aim to synthesize images depicting plausible ways of a human hand interacting with it. Our key insight is that this multi-modal process follows a coarse-to-fine procedure. For example, a mug can either be held by its handle or body, but once decided, the hand articulation is largely driven by the local geometry of the mug. We operationalize this idea by proposing a two-step stochastic approach as shown in Fig. 2.

Given an object image, we first use LayoutNet to predict plausible spatial arrangement of the object and the hand (Sec 3.2). The LayoutNet predicts hand proxy that abstracts away appearance and explicitly specifies 2D location, size and approaching direction of a grasp. This abstraction allows global reasoning of hand-object relations and also enables users to specify the interactions. Then, given the predicted hand proxy and the object image, we synthesize a plausible appearance of an HOI via a ContentNet (Sec 3.3).

To supervise the system, we need pixel-aligned pairs of HOI images and object-only images that depict the exact same objects from the exact same viewpoints with the exact same lighting. We obtain such pairs by inpainting techniques that remove humans from HOI images. We further propose a novel data augmentation to prevent the trained model from overfitting to the inpainting artifacts (Sec 3.4).

3.1. Preliminary: Diffusion models

Diffusion models are probabilistic models [32, 78] that learn to generate samples from a data distribution $p(x)$ by sequentially transforming samples from a tractable distribution $p(x_T)$ (e.g., Gaussian distribution). There are two processes in diffusion models: 1) a forward noise process $q(x_t|x_{t-1})$ that gradually adds a small amount of noise and degrades clean data samples towards the prior Gaussian distribution; 2) a learnable backward denoising process $p(x_{t-1}|x_t)$ that is trained to remove the added noise. The backward process is implemented as a neural network. During inference, a noise vector $x_T$ is sampled from the Gaussian prior and is sequentially denoised by the learned backward model [79, 80]. The training of a diffusion model can be treated as training a denoising autoencoder for L2 loss [86] at various noise levels, i.e., denoise $x_0$ for different $x_t$ given $t$. We adopt the widely used loss term in Denoising Diffusion Probabilistic Models (DDPM) [32, 79], which reconstructs the added noise that corrupted the input samples. Specifically, we use the notation $L_{DDPM}(x; c)$ to denote a DDPM loss term that performs diffusion over $x$ but is also conditioned on $c$ (that are not diffused or denoised):

$$L_{DDPM}(x; c) = \mathbb{E}_{(x, c) \sim \mathcal{N}(0, I), t} \|x - D_\theta(x_t, t, c)\|_2^2,$$

where $x_t$ is a linear combination of the data $x$ and noise $c$, and $D_\theta$ is a denoiser model that takes in the noisy data $x_t$, time $t$ and condition $c$. This also covers the unconditional case as we can simply set $c$ as some null token like $\emptyset$ [33].
3.2. LayoutNet: predicting where to grasp

Given an object image \(I^{obj}\), the LayoutNet aims to generate a plausible HOI layout \(l\) from the learned distribution \(p(l | I^{obj})\). We follow the diffusion model regime that sequentially denoises a noisy layout parameter to output the final layout. For every denoising step, the LayoutNet takes in the (noisy) layout parameter along with the object image and denoises it sequentially, i.e. \(l_{t−1} \sim \phi(l_{t−1} | l_t, I^{obj})\). We splat the layout parameter onto the image space to better reason about 2D spatial relationships to the object image and we further introduce an auxiliary loss term to train diffusion models in the layout parameter space.

Layout parameterization. Hands in HOI images typically appear as hands (from wrist to fingers) with forearms. Based on this observation, we introduce an articulation-agnostic hand proxy that only preserves this basic hand structure. As shown in Fig 2, the layout parameter consists of hand palm size \(a^2\), location \(x, y\) and approaching direction \(\arctan(b_1, b_2)\), i.e. \(l := (a, x, y, b_1, b_2)\). The ratio of hand palm size and forearm width \(s\) remains a constant that is set to the mean value over the training set. We obtain the ground truth parameters from hand detection (for location and size) and hand/forearm segmentation (for orientation).

Predicting Layout. The diffusion-based LayoutNet takes in a noisy 5-parameter vector \(l_0\) with the object image and outputs the denoised layout vector \(l_{T−1}\) (we define \(l_0 = l\)). To better reason about the spatial relation between hand and object, we splat the layout parameter into the image space \(M(l)\). The splatted layout mask is then concatenated with the object image and is passed to the diffusion-based LayoutNet. We splat the layout parameter to 2D by the spatial transformer network [36] that transforms a canonical mask template by a similarity transformation.

DDPM loss for layout. One could directly train the LayoutNet with the DDPM loss (Eq. 1) in the layout parameter space: \(\mathcal{L}_{\text{para}} := \mathcal{L}_{\text{DDPM}}(l, I^{obj})\). However, when diffusing in such a space, multiple parameters can induce an identical layout, such as a size parameter with opposite signs or approaching directions that are scaled by a constant. DDPM loss in the parameter space would penalize predictions even if they guide the parameter to a equivalent one that induce the same layout masks as the ground truth. As the downstream ContentNet only takes in the splatted masks and not their parameters, we propose to directly apply the DDPM loss in the splatted image space (see appendix for details):

\[
\mathcal{L}_{\text{mask}} = \mathbb{E}_{(l_0, I^{obj}), e \sim N(0, I), t} \| M(l_0) - M(l_0) \|_2^2;
\]

where \(l_0 := D_\theta(l_t, t, I^{obj})\) is the output of our trained denoiser that takes in the current noisy layout \(l_t\), the time \(t\) and the object image \(I^{obj}\) for conditioning.

In practice, we apply losses in both the parameter space and image spaces \(\mathcal{L}_{\text{mask}} + \lambda \mathcal{L}_{\text{para}}\) because when the layout parameters are very noisy in the early diffusion steps, the splatted loss in 2D alone is a too-weak training signal.

Network architecture. We implement the backbone network as a UNet with cross-attention layers and initialize it from the pretrained diffusion model [58]. The model takes in images with seven channels as shown in Fig 2: 3 for the object image, 1 for the splatted layout mask and another 3 that blends the layout mask with object image. The noisy layout parameter attends spatially to the feature grid from the UNet’s bottleneck and spits out the denoised layout.

Guided layout generation. The LayoutNet is trained to be conditioned on an object image only but the generation can be guided with additional conditions at test time without retraining. For example, we can condition the network to generate layouts such that their locations are at certain places i.e. \(l \sim p(l_0 | I^{obj}), x = x_0, y = y_0\). We use techniques [80] in diffusion models that hijack the conditions after each diffusion steps with corresponding noise levels. This guided diffusion enables user editing and HOI synthesis for scenes with a consistent hand scale (Sec. 4.3). Please refer to the appendix for LayoutNet implementation details.

3.3. ContentNet: predicting how to grasp

Given the sampled layout \(l\) and the object image \(I^{obj}\), the ContentNet synthesizes a HOI image \(I^{hoi}\). While the synthesized HOI images should respect the provided layout, the generation is still stochastic because hand appearance may vary in shape, finger articulation, skin colors, etc. We leverage the recent success of diffusion models in image synthesis and formulate the articulation network as a image-conditioned diffusion model. As shown in Fig 2, at each step of diffusion, the network takes in channel-wise concatenation of the noisy HOI image, the object image and the splatted mask from the layout parameter and outputs the denoised HOI images \(D_\phi(I^{hoi}, t, I^{obj}, M(l))\).

We implement the image-conditioned diffusion model in the latent space [67,77,85] and finetune it from the inpainting model that is pre-trained on large-scale data. The pre-training is beneficial as the model has learned the prior of retaining the pixels in unmask region and hallucinate to fill the masked region. During finetuning, the model further learns to respect the predicted layout, i.e., retaining the object appearance if not occluded by hand and synthesizing hand and forearm appearance depicting finger articulation, wrist orientation, etc.

3.4. Constructing Paired Training Data

To train such a system, we need pairs of object-only images and HOI image. These pairs need to be pixel-aligned except for the hand regions. One possible way is to use synthetic data [10,28] and render their 3D HOI scene with and without hands. But this introduces domain gap between simulation and the real-world thus hurts generalization. We
Figure 3. Paired Data Generation: Given an HOI image, we first segment out hand (b) and remove it by inpainting (c). Then we use SDEdit [51] to reduce inpainting artifact (d). As inpainting introduces discrepancy between mask and unmasked region (f) while SDEdit undesirably modifies the unmasked object region, we mix up both object images in training.

Data Augmentation. Although the inpainting generates Data Augmentation, with people filtered out therefore it is suitable for our task. Behind hands. The inpainter is trained on millions of data crops from egocentric videos with 80% square padding. Then we segment the hand regions to be removed and pass them to the inpainting system [58] to hallucinate the objects behind hands. The inpainter is trained on millions of data with people filtered out therefore it is suitable for our task.

Datasets Instead of testing with inpainted object images, we evaluate our model on the real object-only images cropped from the frames without hands. The goal is to prevent models from cheating by overfitting to the inpainting artifacts, as justified in the ablations below.

The HOI4D dataset is an egocentric video dataset recording humans in a lab environment interacting with various objects such as kettles, bottles, laptops, etc. The dataset provides manual annotations of hand and object masks, action labels, object categories, instance ID, and ground truth 3D hand poses. We train and evaluate on 10 categories where full annotations are released. For each category, we hold out 5 object instances for evaluation. In total, we collect 126 testing images.

The EPIC-KITCHEN dataset displays more diverse and cluttered scenes. We construct our test set by randomly selecting 10 frames from each video clip. We detect and crop out objects without hands [87]. In total, we collect 500 object-only images for testing.

4.1. Evaluating Image Synthesis

Evaluation Metrics. We evaluate HOI generation using three metrics. First, we report the FID score [31,74], which is widely used for image synthesis that measures the distance between two image sets. We generate 10 samples for every input and calculate FID with 1000 HOI images extracted from the test sets. We further evaluate the physical feasibility of the generated hands by the contact recall metric — it computes the ratio of the generated hands that are in the “in-contact” state by an off-the-shelf hand detector [75]. We also carry out user studies to evaluate their perceptual plausibility. Specifically, we present two images from two randomly selected methods to users and ask them to select the more plausible one. We collect 200 (for HOI4D) and 240 (for EPIC-KITCHEN) answers and report the likelihood of the methods being chosen.

Baselines. We compare our method with three strong image-conditional synthesis baselines. 1) Latent Diffusion Model (LDM) [67] is one of the state-of-the-art generic image generation models that is pre-trained with large-scale image data. We condition the model on the object image and finetune it on HO3Pair dataset. This baseline jointly generates both layout and appearance with one network. 2) Pix2Pix [35] is commonly used for pose-conditioned human/hand synthesis [8,55]. We modify the model to condi-
Figure 4. Visualizing HOI synthesis from our method and three baselines [35,40,67] on HOI4D (left) and EPIC-KITCHEN dataset (right).

Table 1. Quantitative results for HOI synthesis using contact recall, FID score, and a user study on the HOI4D and EPIC-KITCHEN datasets. We compare our method with prior works [35, 40, 67].

<table>
<thead>
<tr>
<th>Method</th>
<th>HOI4D dataset</th>
<th>EPIC-KITCHEN dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contact Recall(%)</td>
<td>FID</td>
</tr>
<tr>
<td>LDM [67]</td>
<td>82.67</td>
<td>72.28</td>
</tr>
<tr>
<td>Pix2Pix [35]</td>
<td>79.50</td>
<td>70.26</td>
</tr>
<tr>
<td>VAE-ContentNet [40]</td>
<td>91.00</td>
<td>78.95</td>
</tr>
<tr>
<td>Ours</td>
<td>91.00</td>
<td>84.21</td>
</tr>
</tbody>
</table>

Results. We visualize the generated HOI images in Fig 4. Pix2Pix typically lacks detailed finger articulation. While LDM and VAE generate more realistic hand articulations than Pix2Pix, the generated hands sometimes do not make contact with the objects. The hand appearance near the contact region is less realistic. In some cases, LDM does not add hands at all to the given object images. In contrast, our model can generate hands with more plausible articulation and the synthesized contact regions are more realistic. This is consistent with the quantitative results in Tab 1. While we perform comparably to the baselines in terms of the FID score, we achieve the best in terms of contact recall. The user study shows that our results are favored the most. This may indicate that humans perceive interaction quality as a more important factor than general image synthesis quality.

Generalizing to EPIC-KITCHEN. Although our model is trained only on the HOI4D dataset with limited scenes and relatively clean backgrounds, our model can generalize to the EPIC-KITCHEN dataset without any finetuning. In Fig 4, the model also generalizes to interact with unseen categories such as scissors and cabinet handles. Tab 1 reports similar trends: performing best in contact recall, comparably well in image synthesis and is favored the most by users.

Ablation: Data Augmentation. Tab 2 shows the benefits of data augmentation to prevent overfitting. Without any data augmentation, the model performs well on the inpainted object images but catastrophically fails on the real ones. When we add aggressive common data augmentations like Gaussian blur and Gaussian noise, the performance improves. Training on SDEdit images further boosts the performance. The results also justify the use of real object images as test set since evaluating on the inpainted object images may not reflect the real performance.

Ablation: LayoutNet Design. We analyze the benefits from our LayoutNet design by reporting contact recall. The LayoutNet predicts more physically feasible hands by taking in the splatted layout masks instead of the 5-parameter layout vector (87.14% vs 78.10%). Moreover, the contact recall drops to 83.96% when the diffusion loss in Sec 3.2 is removed, verifying its contribution to the LayoutNet.
4.2. Evaluating Extracted 3D Hand Poses

Thanks to the realism of the generated HOI images, 3D hand poses can be directly extracted from them by an off-the-shelf hand pose estimator [69]. We conduct a user study to compare the 3D poses extracted from our HOI images against methods that directly predict 3D pose from object images. We present the rendered hand meshes overlaid on the object images to users and are asked to select the more plausible one. In total, we collected 400 and 380 answers from users for HOI4D and EPIC-KITCHEN, respectively.

**Baselines.** While most 3D hand pose generation works require 3D object meshes as inputs, a recent work by Corona et al. (GANHand) [10] can hallucinate hand poses from an object image. Specifically, they first map the object image to a grasp type [14] with the predefined coarse pose and then regress a refinement on top. We finetune their released model on the HO3Pairs datasets with the ground truth 3D hand poses. We additionally implement a diffusion model baseline that sequentially diffuses 3D hand poses. The architecture is mostly the same as the LayoutNet but the diffused parameter is increased to 51 (48 for hand poses and 3 for scale and location) and the splatting function is replaced by the MANO [68] layer that renders hand poses to image. See the appendix for implementation details.

<table>
<thead>
<tr>
<th>Method</th>
<th>HOI4D</th>
<th>EPIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GANHand [10]</td>
<td>23.8</td>
<td>23.53</td>
</tr>
<tr>
<td>3D Pose Diffusion</td>
<td>27.9</td>
<td>34.1</td>
</tr>
<tr>
<td>Ours</td>
<td>48.2</td>
<td>42.4</td>
</tr>
</tbody>
</table>

**Results.** As shown in Fig 5, GANHand [10] predicts reasonable hand poses for some objects but fails when the grasp type is not correctly classified. The hand pose diffusion model sometimes generates infeasible hand poses like acute joint angles. Our model is able to generate hand poses that are compatible with the objects. Furthermore, while previous methods typically assume right hands only, our model can automatically generate both left and right hands by implicitly learning the correlation between approaching direction and hand sides. The qualitative performance is also supported by the user study in Tab 3.

4.3. Application

We showcase several applications that are enabled by the proposed method for hand-object-image synthesis.

**Few-shot Adaptation.** In Tab 4, we show that our model can be quickly adapted to a new HOI category with as few as 32 training samples. We initialize both LayoutNet and ContentNet from our HOI4D-pretrained checkpoints and compare it with the baseline model that was pre-trained for inpainting on a large-scale image dataset [67]. We finetune both models on 32 samples from three novel categories in HOI4D and test with novel instances. The baseline model adapts quickly on some classes, justifying our reasons to finetune our model from them—generic large-scale image pretraining indeed already learns good priors of HOI.
thermore, our HOI synthesis model performs even better than the baseline.

**Layout Editing.** The layout representation allows users to edit and control the generated hand’s structure. As shown in Fig 6, while we gradually change the layout’s location and orientation, the synthesized hand’s appearance changes accordingly. As the approaching direction to the mug changes from right to left, the synthesized fingers change accordingly from pinching around the handle to a wider grip around the mug’s body.

**Heatmap-Guided Synthesis.** As shown in Sec 3.2, our synthesized HOI images can be conditioned on a specified location without any retraining. This not only allows users to edit with just keypoints, but also enables our model to utilize contact heatmap predictions from prior works [13, 56]. In Fig 7, we sample points from the heatmaps and conditionally generate layouts and HOI images which further specifies how to interact at the sampled location.

**Integration to scene.** We integrate our object-centric HOI synthesis to scene-level affordance prediction. While the layout size is predicted relative to each object, hands for different objects in one scene should exhibit consistent scale. To do so, we first specify one shared hand size for each scene and calculate the corresponding relative sizes in each crops (we assume objects at similar depth and thus sizes can be transformed by crop sizes, although more comprehensive view conversions can be used). The LayoutNet is conditioned to generate these specified sizes with guided generation techniques (Sec 3.2). Fig 8 shows the extracted hand meshes from each crops transferred back to the scene.

5. Conclusion

In this paper, we propose to synthesize hand-object interactions from a given object image. We explicitly reason about where to interact and how to interact by LayoutNet and ContentNet. Both of them are implemented as diffusion models to achieve controllable and high-quality visual results. The synthesized HOI images enable a shortcut to more plausible 3D affordance via reconstructing hand poses from them. Although the generation quality and the consistency between the extracted 3D poses and images can be further improved, we believe that HOI synthesis along with our proposed solution opens doors for many promising applications and contributes towards the general goal of understanding human interactions in the wild.

**Acknowledgements.** This work is supported by NVIDIA Graduate Fellowship to YY. The authors would like to thank Xiaolong Wang, Sudeep Dasari, Zekun Hao and Songwei Ge for helpful discussions.
References


[34] Yifei Huang, Minjie Cai, Zhongqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In ECCV, 2018. 2


[80] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3, 4


[92] Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. Learning object placement by inpainting for compositional data augmentation. In *ECCV*, 2020. 5

