Affordance Diffusion: Synthesizing Hand-Object Interactions

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

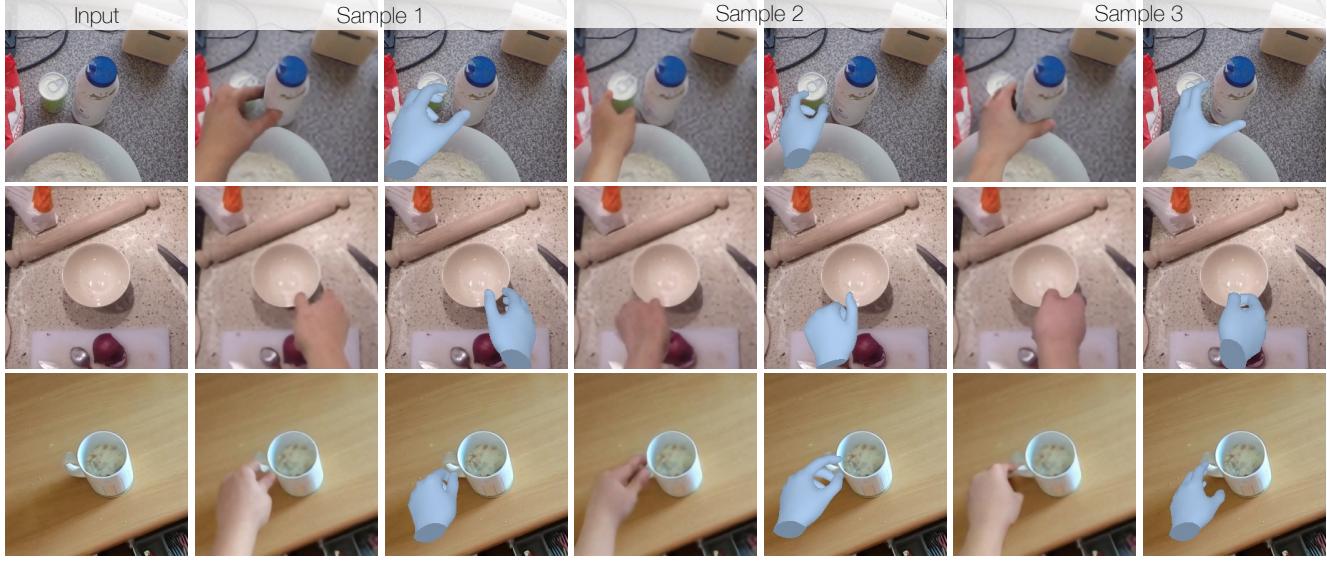https://judyye.github.io/affdiffusion-www

Figure 1. Given a single RGB image of an object (first column), we synthesize plausible images of hand-object interactions from which feasible 3D hand poses can be directly extracted (remaining columns).

Abstract

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.

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 [44, 74] from human faces to ImageNet classes. Newer diffusion models such as Dall-E 2 [66] and Stable Diffusion [67] 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" [66,67].

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

---

*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 [20]. 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 *where* to interact (*layout*) from *how* to interact (*content*) [26, 31]. 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 [12, 49]. Our method 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 [14, 57] 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 techinuqes 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, 54, 76] or segmentation [16, 77]. Many works reconstruct the underlying shape of hands and objects from RGB(D) images or videos by either template-based [6, 19, 27, 83, 85] or template-free methods [10, 29, 39, 91]. Furthermore, temporal understanding of HOI videos [21, 28, 63, 64, 82] 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 [20]. 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, 31, 45, 58]. Some other approaches ground these action labels to images by predicting heatmaps that indicate interaction possibilities [14, 35, 48, 57, 60]. While heatmaps only specify *where* to interact without telling *what* to do, recent approaches predict richer properties such as contact distance [39], action trajectory [48, 55], grasping categories [23, 50], *etc*. Instead of predicting more sophisticated interaction states, we explore directly synthesizing HOI images for possible interactions because images demonstrate both *where* and *how* to interact comprehensively and in a straightforward manner.

3D Affordance. While 3D prediction from images is common for human-scene interaction [17, 24, 26, 46, 94, 95], 3D affordance for hand and object typically takes in *known* object 3D shapes and predicts *stable* grasps [25, 38, 53, 61]. In contrast, our work predicts *coarse* 3D grasps from RGB images of *general* objects. The coarse but generalizable 3D hand pose prediction is shown as useful human prior for dexterous manipulation [2, 13, 42, 50, 65, 89]. While some recent works [11, 39] 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 [33, 80] have driven significant advances in various domains [43, 84, 90, 92, 96], including image synthesis [3, 66, 67, 73]. A key advantage of diffusion models over other families of generative models [22, 41] is their ability to easily adapt to image editing and re-synthesis tasks without much training [18, 40, 71]. While recent image-conditioned generative models achieve impressive results on various image translation tasks such as image editing [5, 36, 51],
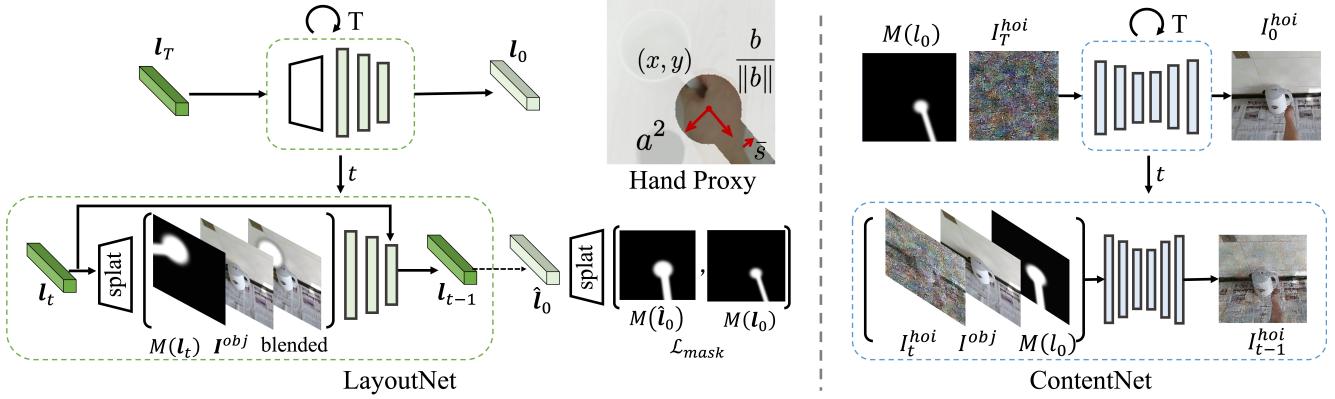
Figure 2. The proposed method consists of two image-conditioned diffusion models: LayoutNet and ContentNet. Given an object image, we first use LayoutNet (left) to predict a HOI spatial arrangement l_0. For every diffusion step, the LayoutNet splats the noisy layout parameter into image space, concatenates it with the object image and their blending, and predicts the denoised layout. We apply the diffusion loss in the splatted 2D space $\mathcal{L}_{\text{mask}}$. Then the ContentNet (right) takes in the predicted layout along with the object image to synthesize an HOI image. The two modules are connected by the articulation-agnostic hand proxy (middle top).

style transfer [47, 62], the edits mostly modify textures and style, but preserve structures, or insert new content to user-specified regions [1, 66, 72]. In contrast, we focus on affordance synthesis where both layout (structure) and appearance are automatically reasoned about.

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 a 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). This allows the network to implicitly reason about 3D wrist orientation, finger placement, and occlusion based on the object's local shape. We use conditional diffusion models for both networks to achieve high-quality layout and visual content. The synthesized HOI image is realistic such that a feasible 3D hand pose can be directly extracted from it by an off-the-shelf hand pose reconstruction model (Sec 4.2).

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 ex-

act 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 [33, 79] that learn to generate samples from a data distribution $p(\mathbf{x})$ by sequentially transforming samples from a tractable distribution $p(\mathbf{x}_T)$ (*e.g.*, Gaussian distribution). There are two processes in diffusion models: 1) a forward noise process $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ that gradually adds a small amount of noise and degrades clean data samples towards the prior Gaussian distribution; 2) a learnable backward denoising process $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$ that is trained to remove the added noise. The backward process is implemented as a neural network. During inference, a noise vector $\mathbf{x}_T$ is sampled from the Gaussian prior and is sequentially denoised by the learned backward model [80, 81]. The training of a diffusion model can be treated as training a denoising autoencoder for L2 loss [87] at various noise levels, *i.e.*, denoise $\mathbf{x}_0$ for different $\mathbf{x}_t$ given t. We adopt the widely used loss term in Denoising Diffusion Probabilistic Models (DDPM) [33, 80], which reconstructs the added noise that corrupted the input samples. Specifically, we use the notation $\mathcal{L}_{\text{DDPM}}[\mathbf{x}; \mathbf{c}]$ to denote a DDPM loss term that performs diffusion over $\mathbf{x}$ but is also conditioned on $\mathbf{c}$ (that are not diffused or denoised):

$$\mathcal{L}_{\text{DDPM}}[\mathbf{x}; \mathbf{c}] = \mathbb{E}_{(\mathbf{x},\mathbf{c}),\epsilon\sim\mathcal{N}(0,I),t}\|\mathbf{x} - D_\theta(\mathbf{x}_t, t, \mathbf{c})\|_2^2, \quad (1)$$

where $\mathbf{x}_t$ is a linear combination of the data $\mathbf{x}$ and noise ϵ, and D_θ is a denoiser model that takes in the noisy data $\mathbf{x}_t$, time t and condition $\mathbf{c}$. This also covers the unconditional case as we can simply set $\mathbf{c}$ as some null token like $\varnothing$ [34].

3.2. LayoutNet: predicting where to grasp

Given an object image $\mathbf{I}^{obj}$, the LayoutNet aims to generate a plausible HOI layout l from the learned distribution $p(l|\mathbf{I}^{obj})$. 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.* $l_{t-1} \sim \phi(l_{t-1}|l_t, \mathbf{I}^{obj})$. 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 $\bar{s}$ 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_t 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 $M(l_t)$. 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 [37] 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: $\mathcal{L}_{para} := \mathcal{L}_{\text{DDPM}}[l; \mathbf{I}^{obj}]$. 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):

$$\mathcal{L}_{mask} = \mathbb{E}_{(l_0, \mathbf{I}^{obj}), \epsilon \sim \mathcal{N}(0, I), t} \| M(l_0) - M(\hat{l}_0) \|_2^2. \quad (2)$$

where $\hat{l}_0 := D_\theta(l_t, t, \mathbf{I}^{obj})$ is the output of our trained denoiser that takes in the current noisy layout l_t, the time t and the object image $\mathbf{I}^{obj}$ for conditioning.

In practice, we apply losses in both the parameter space and image spaces $\mathcal{L}_{mask} + \lambda \mathcal{L}_{para}$ 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 [59]. 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 spit out the denoised output.

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.* $l \sim p(l_0|\mathbf{I}^{obj}, x = x_0, y = y_0)$. We use techniques [81] 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.
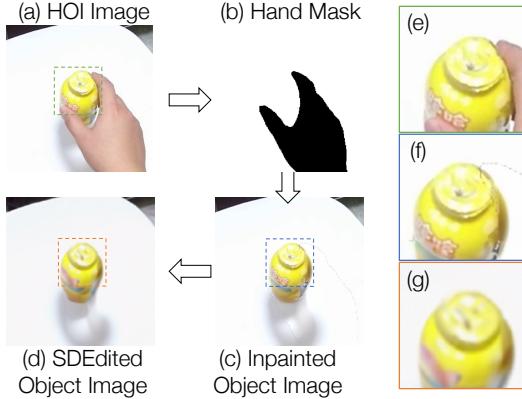
3.3. ContentNet: predicting how to grasp

Given the sampled layout l and the object image $\mathbf{I}^{obj}$, the ContentNet synthesizes a HOI image $\mathbf{I}^{hoi}$. 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 $D_\phi(\mathbf{I}_t^{hoi}, t, [\mathbf{I}^{obj}, M(l)])$.

We implement the image-conditioned diffusion model in the latent space [68, 78, 86] and finetune it from the inpainting model that is pre-trained on large-scale data. The pretraining 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 [11, 29] 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

(a) HOI Image   (b) Hand Mask   (e)

(f)

(g)

(d) SDEdited   (c) Inpainted
Object Image   Object Image

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 [52] to reduce inpainting artifact (d). As inpainting introduce discrepancy between mask and unmasked region (f) while SDEdit undesirably modifies the unmasked object region, we mix up *both* object image sets in training.

instead follow a different approach.

As shown in Fig 3, we first extract object-centric HOI crops from egocentric videos with 80% square padding. Then we segment the hand regions to be removed and pass them to the inpainting system [59] 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.

Data Augmentation. Although the inpainting generates impressive object-only images, it still introduces editing artifacts, which the networks can easily overfit to [93], such as sharp boundary and blurriness in masked regions. We use SDEdit [51] to reduce the discrepancy between the masked and unmasked regions. SDEdit first adds a small amount of noise (we use 5% of the whole diffusion process) to the given image and then denoises it to optimize overall image realism. However, although the discrepancy within images reduces, the unmasked object region is undesirably modified and the overall SDEdited images appear blurrier. In practice, we mix up the object-only images with and without SDEdit for training.

We collect all data pairs from HOI4D [49]. After some automatic sanity filtering (such as ensuring hands are removed), we generate 364k pairs of object-only images and HOI-images in total. We call the dataset HO3Pairs (Hand-Object interaction and Object-Only Pairs). We provide details and more examples of the dataset in the appendix.

4. Experiments

We train our model on the contructed HO3Pairs dataset, evaluate it on the HOI4D [49] dataset and show *zero-shot* generalization to the EPIC-KITCHEN [12] dataset. We evaluate both the generated HOI images and the extracted 3D poses. For image synthesis, we compare with condi-

tional image synthesis baselines and show that our method generates more plausible hands in interaction. Beyond 2D HOI image synthesis, we compare the extracted 3D poses with prior works that directly predict 3D hand poses. Furthermore, we show several applications enabled by the proposed HOI synthesis method, including few-shot adaptation, image editing by layout, heatmap-guided prediction and integrating object affordance with the scene.

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 [88]. 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 [32,75], 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 [76]. 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)* [68] 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* [36] is commonly used for pose-conditioned human/hand synthesis [9,56]. We modify the model to condi-
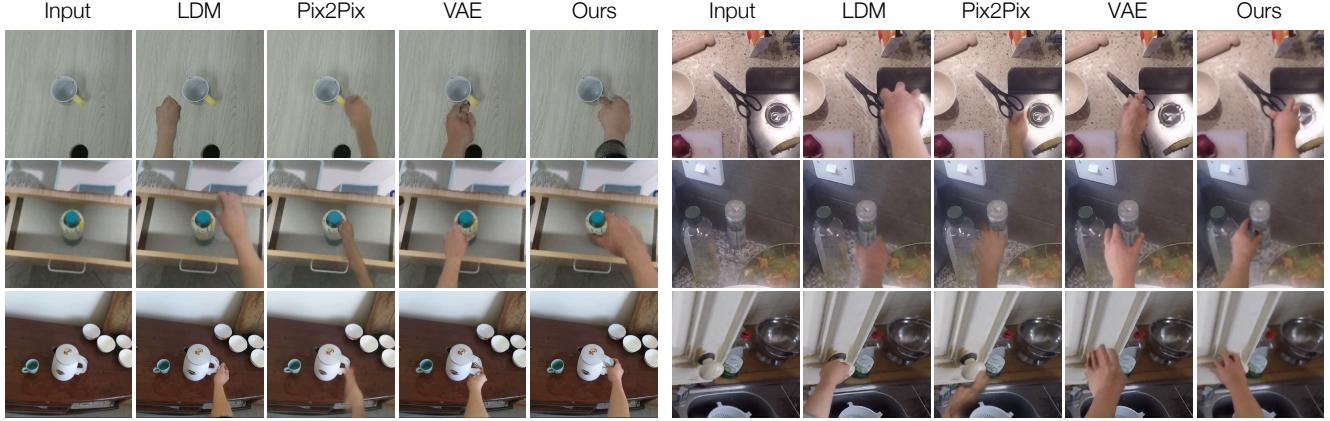
Figure 4. Visualizing HOI synthesis from our method and three baselines [36,41,68] on HOI4D (left) and EPIC-KITCHEN dataset (right).

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 [36,41,68].

| Method | HOI4D dataset | | | | | | | | | | | | EPIC-KITCHEN dataset | | |
| | Contact Recall(%) | | | | | | | | | | FID | User Study | Contact Recall | FID | User Study |
	Kettle	Knife	TrashCan	Chair	Mug	Bowl	ToyCar	Laptop	Bottle	mean					
LDM [68]	82.67	72.28	83.33	82.08	66.67	78.10	88.00	62.00	87.22	64.44	105.26	27.5	76.56	118.15	23.3
Pix2Pix [36]	79.50	70.26	82.50	76.88	68.50	79.64	89.00	63.00	85.42	73.02	107.09	15.5	70.00	125.62	13.3
VAE-ContentNet [41]	91.00	78.95	91.50	85.63	73.00	90.00	94.00	69.00	90.00	83.49	**98.19**	23.0	82.03	**115.86**	27.9
Ours	91.00	84.21	97.00	88.75	60.00	92.86	96.00	72.00	91.67	**87.14**	99.00	**34.0**	**86.56**	117.22	**35.4**

tion on the generated layout masks that are predicted from our LayoutNet. 3) *VAE* [41] is a widely applied generative model in recent affordance literature [17,46,94]. This baseline uses a VAE with ResNet [30] as backbone to predict a layout parameter. The layout is then passed to our Content-Net to generate images.

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, compara-

Table 2. **Analysis of data augmentation**: contact recall (CR%) and FID score on the real and the inpainted object image set of HOI4D and comparisons of ours with the ablations of excluding aggressive common data augmentation (CmnAug) or SDEdit [52].

| | | Real Obj Img | | Inpainted Img | |
CmnAug	SDEdit	CR	FID	CR	FID
		39.37	113.93	89.05	89.38
✓		79.52	99.12	93.81	89.01
✓	✓	87.14	99.00	94.29	88.50

bly 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 SDEdited 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.
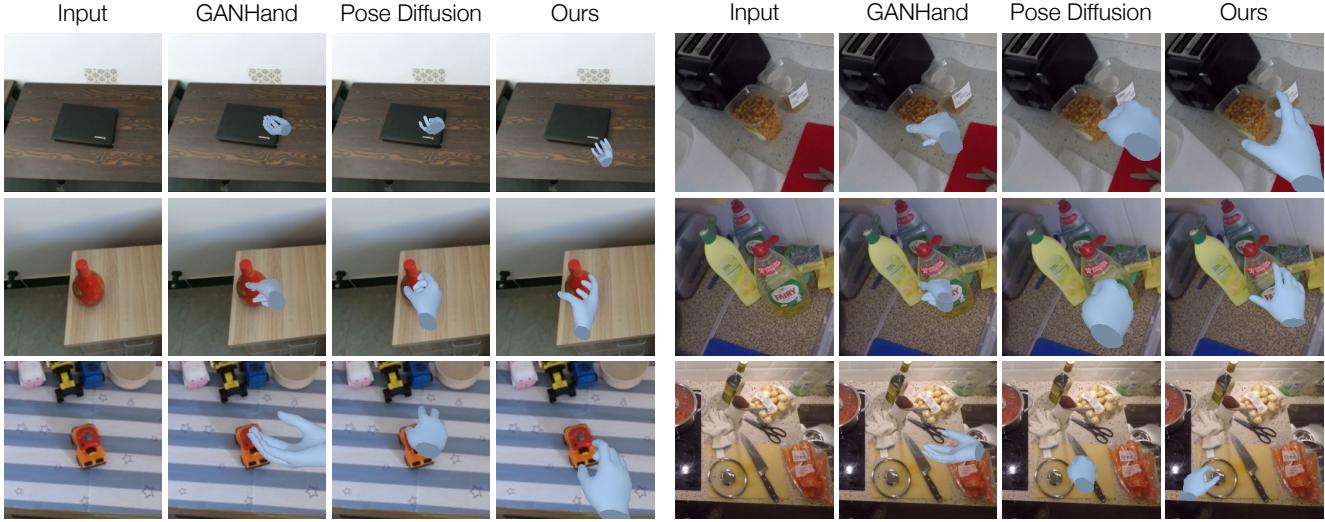
6

Figure 5. Visualizing 3D affordance prediction from our method, GANHand [11] and diffusion model [68] that directly predicts 3D pose on HOI4D (left) and EPIC-KITCHEN dataset (right).

Table 3. User study for 3D affordance prediction on HOI4D and EPIC-KITCHEN dataset. We compare our method with GAN-Hand [11] and a diffusion model that directly predicts 3D poses.

Method	HOI4D	EPIC
GANHand [11]	23.8	23.53
3D Pose Diffusion	27.9	34.1
Ours	**48.2**	**42.4**

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 [70]. 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) [11] can hallucinate hand poses from an object image. Specifically, they first map the object image to a grasp type [15] 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 [69] layer that renders hand poses to image.

Table 4. **Few-shot Adaption:** Quantitative results using contact recall when finetuning the proposed HOI synthesis model and a pretrained inpainting model with 32 samples from new categories.

	bucket	scissors	stapler	mean
w HOI pretrain	92.0	95.0	70.0	85.7
w/o HOI pretrain	90.0	68.8	34.0	64.3

See the appendix for implementation details.

Results. As shown in Fig 5, GANHand [11] 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 [68]. 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
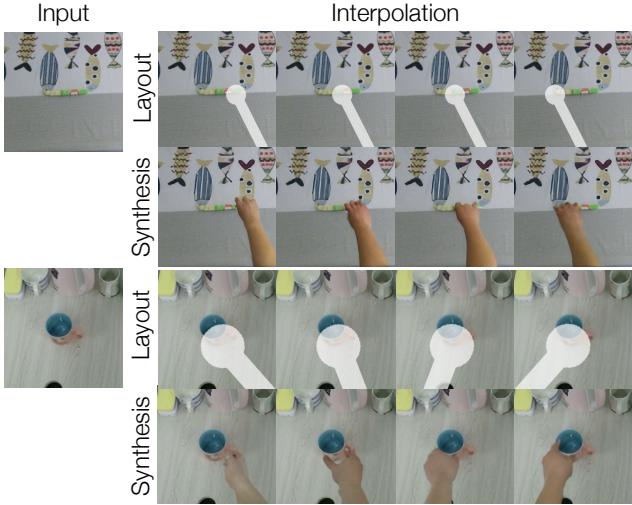
Figure 6. **Layout Editing**: Visualizing HOI synthesis when the conditioned layouts gradually change location and orientation.
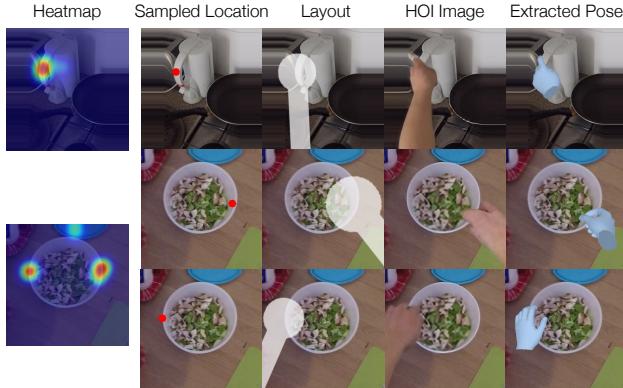

Figure 7. **Heatmap-guided synthesis:** Given a heatmap, LayoutNet is guided to generate layout at the sampled location, from which HOI images are synthesized and 3D poses are extracted.

pretraining indeed already learns good priors of HOI. Furthermore, 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 [14, 57].

Figure 8. **Scene-level Integration:** Given a cluttered scene, we detect each object and synthesize its interactions individually. Each object's layout scale is guided to appear in the same size when transferred back to the scene.

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. Conslusion

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.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 3

[2] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *RSS*, 2022. 2

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, 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*, 2022. 2

[4] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015. 2

[5] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *ECCV*, 2022. 2

[6] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *ECCV*, 2020. 2

[7] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, 2016. 2

[8] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. *arXiv*, 2015. 13

[9] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019. 5

[10] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *CVPR*, 2022. 2

[11] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020. 2, 4, 7, 13, 14, 19

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 2, 5

[13] Sudeep Dasari, Abhinav Gupta, and Vikash Kumar. Learning dexterous manipulation from exemplar object trajectories and pre-grasps. In *ICRA*, 2023. 2

[14] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, 2018. 2, 8

[15] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on Human-Machine Systems*, 2015. 7

[16] Victoria Florence, Jason J Corso, and Brent Griffin. Robot-supervised learning for object segmentation. In *ICRA*, 2020. 2

[17] David F Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A Efros, Ivan Laptev, and Josef Sivic. People watching: Human actions as a cue for single view geometry. In *ECCV*, 2012. 2, 6

[18] 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*, 2022. 2

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

[20] James J Gibson. The ecological approach to the visual perception of pictures. *Leonardo*, 1978. 2

[21] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *ICCV*, 2021. 2

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 2

[23] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *CVPR*, 2022. 2

[24] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *CVPR*, 2011. 2

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

[26] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011. 2

[27] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *CVPR*, 2010. 2

[28] Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. Tracking a hand manipulating an object. In *ICCV*, 2009. 2

[29] 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 *CVPR*, 2019. 2, 4

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[31] Tucker Hermans, James M Rehg, and Aaron Bobick. Affordance prediction via learned object attributes. In *ICRA Workshop*, 2011. 2

[32] 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. *NeurIPS*, 2017. 5

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2, 3

[34] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NerIPS Workshop*, 2022. 3

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

[36] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2, 5, 6, 13, 14, 17, 18

[37] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *NeurIPS*, 2015. 4, 12

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

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

[40] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv*, 2022. 2

[41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014. 2, 6, 13, 14, 17, 18

[42] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning task-oriented grasping from human activity datasets. *IEEE Robotics and Automation Letters*, 2020. 2

[43] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *ICLR*, 2021. 2

[44] Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *CVPR*, 2022. 1

[45] Yong Jae Lee and Kristen Grauman. Predicting important objects for egocentric video summarization. *IJCV*, 2015. 2

[46] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *CVPR*, 2019. 2, 6

[47] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *CoRR*, 2022. 3

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

[49] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *CVPR*, 2022. 2, 5

[50] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *CoRL*, 2022. 2

[51] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2, 5

[52] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *ICLR*, 2022. 5, 6

[53] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *Robot Autom Mag*, 2004. 2

[54] Arpit Mittal, Andrew Zisserman, and Philip HS Torr. Hand detection using multiple proposals. In *BMVC*, 2011.

[55] Kaichun Mo, Leonidas J. Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2Act: From pixels to actions for articulated 3D objects. In *ICCV*, 2021. 2

[56] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. GANerated hands for real-time 3D hand tracking from monocular RGB. In *CVPR*, 2018. 5

[57] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *ICCV*, 2019. 2, 8

[58] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *CVPR*, 2020. 2

[59] 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. *ICML*, 2022. 4, 5, 12, 13

[60] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv*, 2017. 2

[61] Andreas ten Pas and Robert Platt. Using geometry to detect grasps in 3d point clouds. *arXiv*, 2015. 2

[62] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 3

[63] Will Price, Carl Vondrick, and Dima Damen. Unweavenet: Unweaving activity stories. In *CVPR*, 2022. 2

[64] Senthil Purushwalkam, Tian Ye, Saurabh Gupta, and Abhinav Gupta. Aligning videos in space and time. In *ECCV*, 2020. 2

[65] 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 *ECCV*, 2022. 2

[66] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv*, 2022. 1, 2, 3, 12

[67] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 13

[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4, 5, 6, 7, 12, 13, 14, 17, 18, 19

[69] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 2017. 7

[70] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *ICCVW*, 2021. 7

[71] 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. *arXiv*, 2022. 2

[72] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, 2022. 3

[73] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv*, 2022. 2

[74] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH*, 2022. 1

[75] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. 5

[76] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 2, 5, 12

[77] Dandan Shan, Richard Higgins, and David Fouhey. Cohesiv: Contrastive object and hand embedding segmentation in video. *NeurIPS*, 2021. 2

[78] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *NeurIPS*, 2021. 4

[79] Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *ICML*, 2015. 3

[80] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3, 12

[81] 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

[82] Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPRW*, 2009. 2

[83] 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 *ECCV*, 2016. 2

[84] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *NeurIPS*, 2021. 2

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

[86] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *NeurIPS*, 2021. 4

[87] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7), 2011. 3

[88] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[89] Yueh-Hua Wu, Jiashun Wang, and Xiaolong Wang. Learning generalizable dexterous manipulation from human grasp affordance. *CoRL*, 2022. 2

[90] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A geometric diffusion model for molecular conformation generation. In *ICLR*, 2022. 2

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

[92] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In *NeurIPS*, 2022. 2

[93] 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

[94] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *3DV*, 2020. 2, 6

[95] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3D people in scenes without people. In *CVPR*, 2020. 2

[96] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, 2021. 2

Affordance Diffusion: Synthesizing Hand-Object Interactions Supplementary Materials

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

In the supplementary material, we provide more implementation details and more qualitative results. We discuss the details of articulation-agnostic hand proxy and how to apply DDPM loss in the image space for training the LayoutNet (Sec. A.1). We also present ablations on ContentNet(Sec. A.2). We further show: (i) the paired data construction method being robust, in Sec. A.3, (ii) baseline implementations details in Sec. A.4, (iii) details of integrating our approach to scene-level affordance prediction in Sec. A.5. Finally, we discuss the limitation of our approach (Sec. A.6), and show more qualitative results in Sec. B. **Visual results are also included in the video.**

A. Implementation Details

A.1. LayoutNet (Sec 3.1)

Layout parameters. As mentioned in Sec 3.1 of the main paper, we parameterize the layout as (x, y, a, b_1, b_2), where x, y is the location, a^2 is size, and b_1, b_2 are un-normalized approaching direction parameters. For training the LayoutNet, we obtain the ground truth parameters from off-the-shelf 2D hand prediction systems. The size and location comes from the predicted bounding box of a hand detector [76], which typically defines the hand region up to the wrist. The orientation is calculated from hand segmentation whose region is typically defined as the entire hand region, including hand and forearm. The approaching direction is calculated as the first principal component of a hand mask that centers on the location of the palm of the predicted hand.

We splat the layout parameters onto 2D via the spatial transformer network [37] that transforms a canonical mask template by a similarity transformation. The 2D similarity transformation is determined from the layout parameters. More formally,

$$T_l = \begin{pmatrix} sR & t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a^2\hat{b}_1 & -a^2\hat{b}_2 & x \\ a^2\hat{b}_2 & a^2\hat{b}_1 & y \\ 0 & 0 & 1 \end{pmatrix},$$

where $\hat{b}_1, \hat{b}_2$ is the normalized vector of b_1, b_2.

The lollipop-shape template in the canonical space is implemented with its circle being an isometric 2D Gaussian with a standard deviation of 1 and its rectangle being a 1D Gaussian with a standard deviation $\bar{s} = 2$. The width of the rectangle is calculated from the training data as the average ratio of the widths of forearms and palms.

DDPM loss on mask. In Eq 1 and 2 of the main paper, we write the DDPM loss in terms of reconstructing clean samples. In practice, we follow prior works [59, 66, 68] that reconstruct the added noise ϵ as

$$\mathcal{L}_{\text{DDPM}}^{\text{noise}} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,I),t}\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2.$$

The estimated clean sample $\hat{l}_0$ is connected with the estimated noise by $\hat{l}_0 = \frac{1}{\sqrt{1-\bar{\alpha}_t}}l_t - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta$, where $\alpha_t, \bar{\alpha}_t$ represent the noise schedule for each diffusion time step.

We train the LayoutNet with a weighted sum of the parameter loss $\mathcal{L}_{\text{para}}$ for esitmating the noise term ϵ, and a mask loss $\mathcal{L}_{\text{mask}}$ for estimating the clean sample term $\hat{l}_0$. The hyperparamter λ is set to 10.

Guided layout generation. LayoutNet inherits properties from diffusion models that can be guided to generate samples with additional constraints at test time. We follow Song *et al.* [80]. After each diffusion steps, we hijack the additional constraints with corresponding noise levels for the next diffusion step.

More specifically, instead of passing in the network's output x_t from the previous time step, we hijack it with $x_t \leftarrow \tilde{x}_t m + x_t(1 - m)$, where m is the indicator mask of the given condition $\tilde{x}_0$. The unspecified constraints in $\tilde{x}_0$ are set to 0. $\tilde{x}_t$ represents the additional constraint with corresponding noise level, *i.e.* $\sqrt{1 - \bar{\alpha}_t}\tilde{x}_0 + \sqrt{\bar{\alpha}_t}\epsilon$.

A.2. ContentNet (Sec3.2)

The goal of ContentNet is to generate high-resolution (256^2) realistic HOI images conditioned on the predicted layout and the input object image. We tried two different approaches commonly used in diffusion models [59, 68] as backbones for the ContentNet. One way (called ours/AffordDiff-LDM) is to follow Rombach *et al*. [67], as described in our main paper, that implements the ContentNet in the latent space where images of size 256^2 are compressed to 3-dimensional features of size 64^2 by a fixed pretrained autoencoder. The other way (called ours/AffordDiff-GLIDE) is to follow Nichol *et al*. [59] that uses a cascaded diffusion model that first generates images of size 64^2 and then upsamples them by a factor of 4.

All of the quantitative results in our main paper, including the user studies and all ablations, are based on Afford-LDM. AffordDiff-GLIDE is better in terms of contact recall (90.8% vs 87.1%) while AffordDiff-LDM is significantly better in terms of FID score (99.0 vs 121.6). We find that AffordDiff-LDM generates less blurry results and the hand texture appears sharper and more realistic. In comparison, we find AffordDiff-GLIDE perceptually preferred because AffordDiff-GLIDE generates more realistic, though blurrier, finger articulations. The qualitative results in the main paper on EPIC-KITCHEN dataset (Fig 1 and Fig4 right in the main paper) show Afford-GLIDE. However, we provide the qualitative comparison of Afford-LDM with baselines in Fig 9 and Fig 10 of the appendix. We further provide a comparison of these two variants in Fig 15 of the appendix.

A.3. Constructing Paired Training Data (Sec3.3)

Cropping Details. We crop all objects with 80% squared padding before resizing such that objects (hands) appear in similar (different) sizes. The model learns the priors of their relative scales, *e.g*., a hand to grasp a kettle appears much smaller than that of a mug (Fig 4).

We show that the proposed method to obtain pixel-aligned pairs of HOI and object-only images is robust and can also be applied to more cluttered images. When there is more than one hand in the HOI image, we randomly select one to remove. We show results of applying our data construction method on the HOI4D (Fig 9) and the EPIC-KITCHEN (Fig 10) datasets.

A.4. Baselines Implementation

Pix2Pix [36] (Sec4.1) We modify the official Pix2Pix implementation[1]. Given the predicted layout and the provided object image, we concatenate them channel-wise and pass them through 6 blocks of ResNet to output HOI images. The discriminator takes in the concatenation the of the object-only image, the splatted layout image, and generated

---

[1]https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

HOI image and learns to discriminate between the real and fake domains. We tried batchnorm and instancenorm and found that batchnorm generated better results in general but has some black holes if the background statistics deviate from that of the training set.

VAE [41] (Sec4.1) VAE is notoriously known for being hard to balance for both generation variance and reconstruction quality. We sweep hyperparameters of the KL divergence loss's weights from $1, 1e-1, 1e-2, 1e-3, 1e-4$ and use $1e-3$ as it produces the highest contact recall.

GANHand [11] (Sec4.2) GANHand is originally proposed both to predict 3D MANO hands for images of YCB objects [8] and to optimize physical plausibility with respect to the known or reconstructed 3D shapes of YCB objects. We compare our method with their sub-network for grasp prediction from RGB images (blue branch in their original paper, Fig 4). The sub-network takes in the object's identity, the desk plane equation and the object's center in 3D space, in addition to the object image. Since these are not available in the HOI4D dataset, we set them to zeros. We apply an additional reconstruction loss for 3D hand joints, MANO hand parameters and camera parameters. We fine-tune the network from the public checkpoints for another 10k iterations.

A.5. Scene Integration

We integrate our object-centric HOI synthesis to scene-level affordance prediction. We first detect the objects in the scene and then expand the detected bounding box's size with the same pad ratio (0.8 of the original object size). However, when the scene is crowded, the extended object crops may include other objects thus distracting the layout generation. We instead crop the object with the detected bounding box and pad the cropped object with boundary values. This allows the network to generate hand interaction only for the object of interest.

A.6. Limitation and Failure Cases

Although it is encouraging that the proposed model can perform zero-shot generalization to the EPIC-KITCHEN dataset, the proposed method inherits limited generalization capabilities from general learning-based algorithms. The proposed model will fail when the object image's appearance deviates too much from the training set, *e.g.* for too cluttered scenes, extreme lighting, very large objects (like a fridge) or very small objects (like a pin), *etc*. The current model also cannot generate hands entering from the top of the frame or generate hands from a third-person's view due to the bias in the training set. These limitations require training with more diverse data. Additionally, the consistency of the hand's appearance and of the extracted hand poses can be further improved.
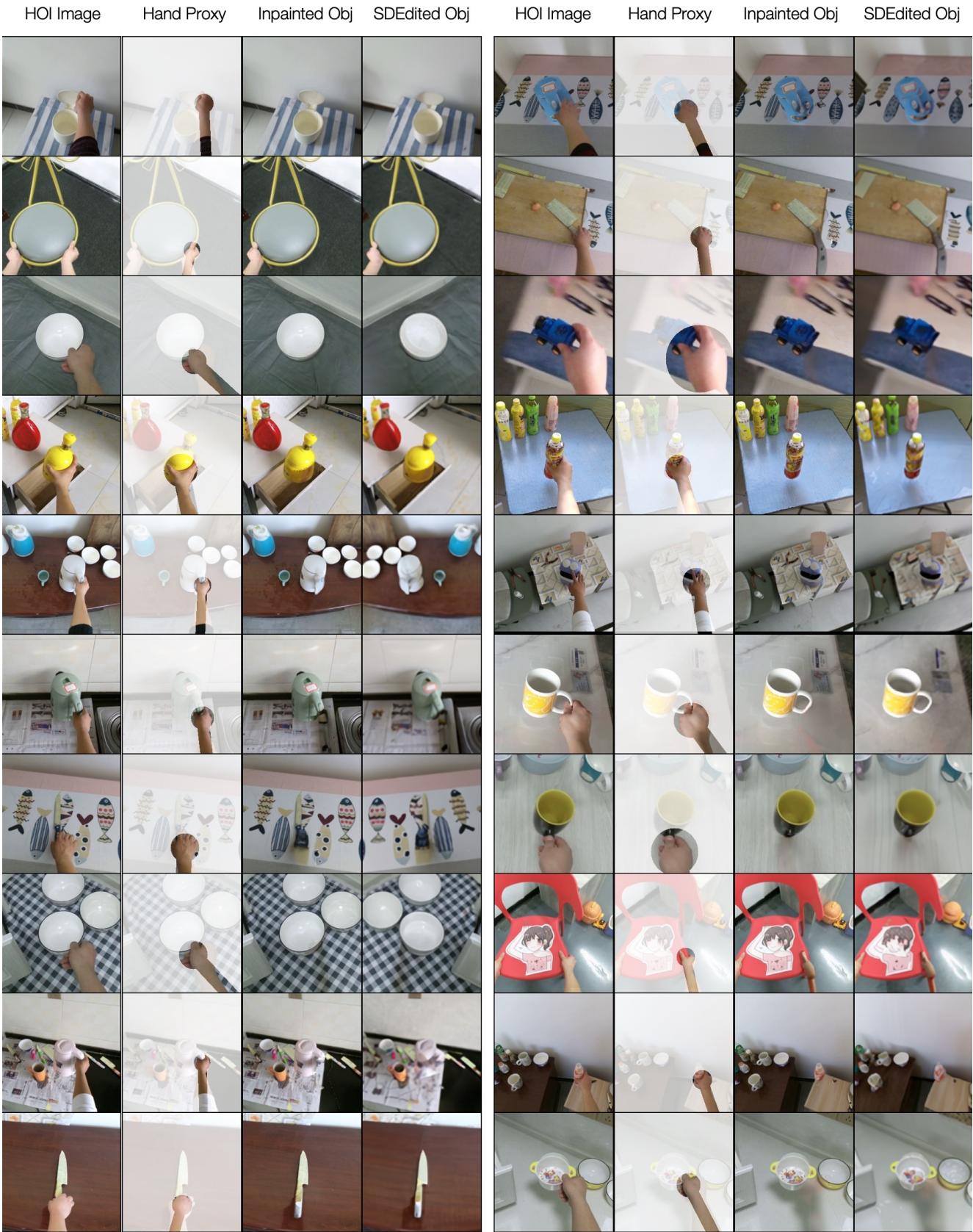
B. Qualitative Results

Fig 9 shows more examples of the constructed paired training data. We train all the models with a uniform mixture of inpainted and SDEdited object images.

Fig 10 shows that the proposed paired data construction is robust and can be applied to the EPIC-KITCHEN dataset.

Fig 11 shows more comparisons of the generated HOI images by the proposed method (LDM-version as reported in tables) and other image synthesis baselines [36,41,68] on the HOI4D dataset.

Fig 12 shows more comparisons of the generated HOI images by the proposed method (LDM-version as reported in tables) and other image synthesis baselines [36,41,68] on EPIC-KITCHEN dataset.

Fig 13 shows more comparisons of the extracted 3D hand pose obtained by the proposed method and other 3D affordance baselines [11,68] on the HOI4D dataset.

Fig 14 shows more comparisons of the extracted 3D hand pose obtained by the proposed method and other 3D affordance baselines [11,68] on the EPIC-KITCHEN dataset.

Fig 15 shows an ablation study on comparison of the LDM and GLIDE version of our model on HOI4D and EPIC-KITCHEN datasets.

Fig 16 shows more layout editing results.

Fig 17 shows more results of heatmap-guided synthesis.

Figure 9. Visualizing more examples of the constructed paired training data. We train all the models with a mixture of inpainted and SDEdited object images.

HOI Image   Hand Proxy   Inpainted Obj   SDEdited Obj   HOI Image   Hand Proxy   Inpainted Obj   SDEdited Obj
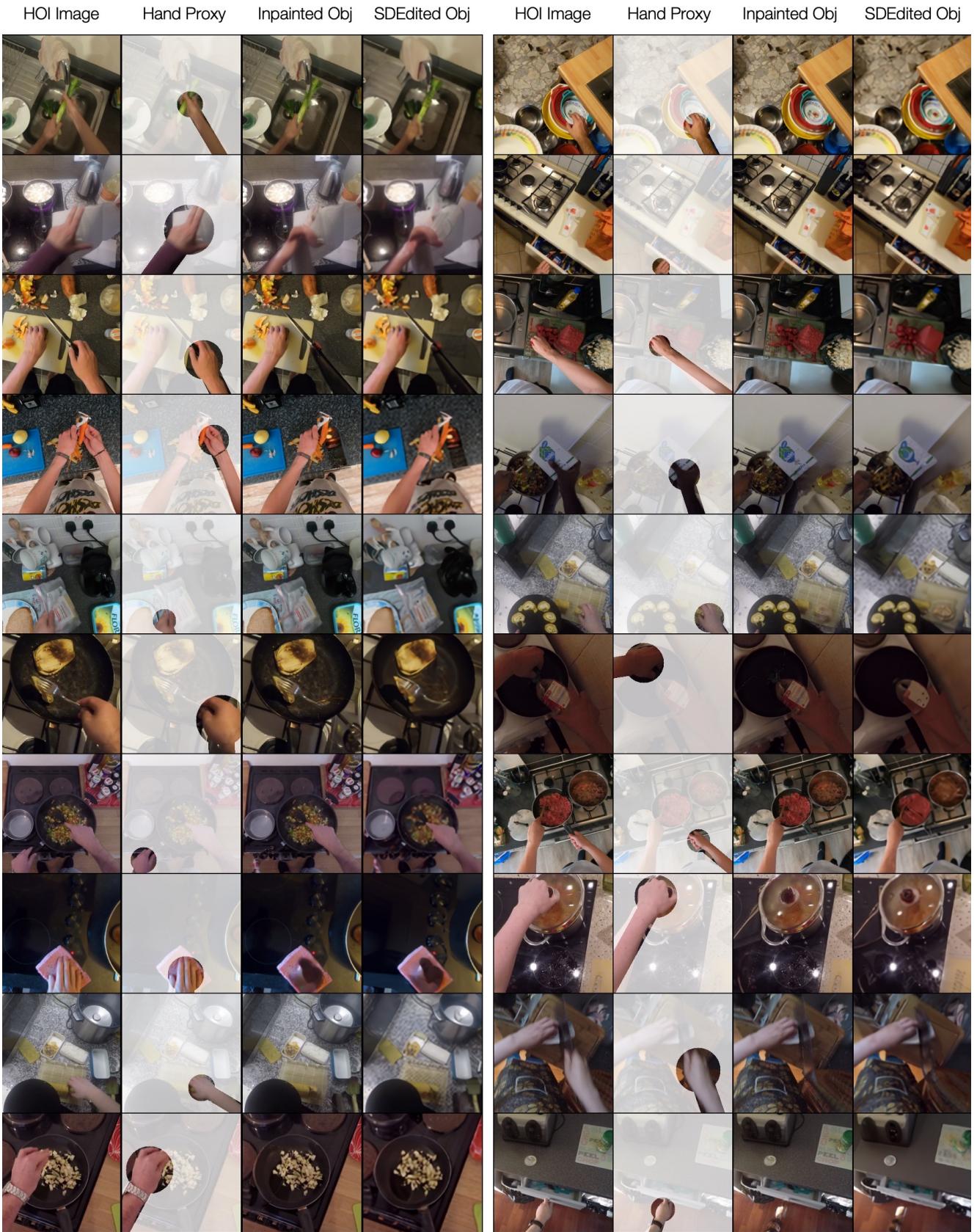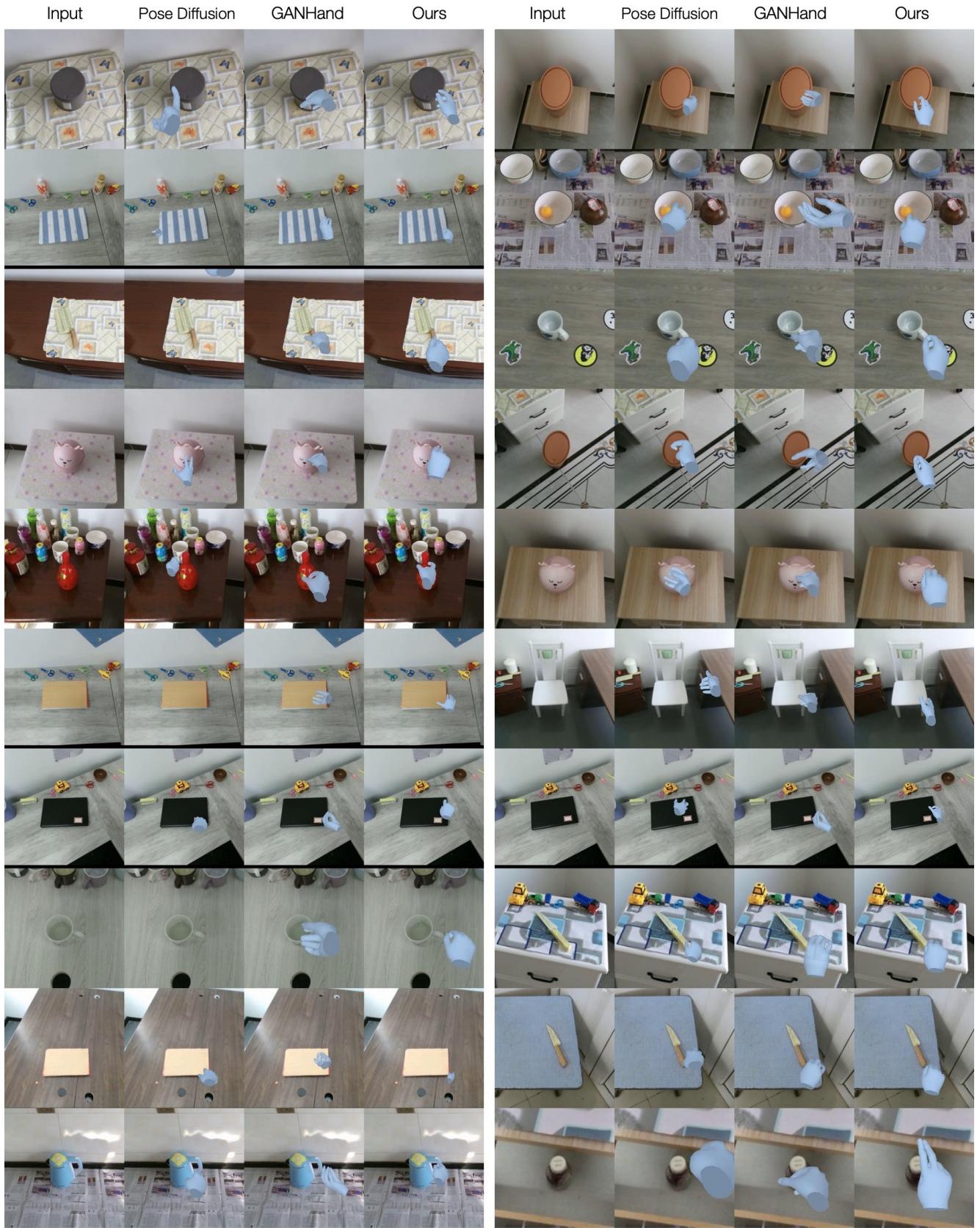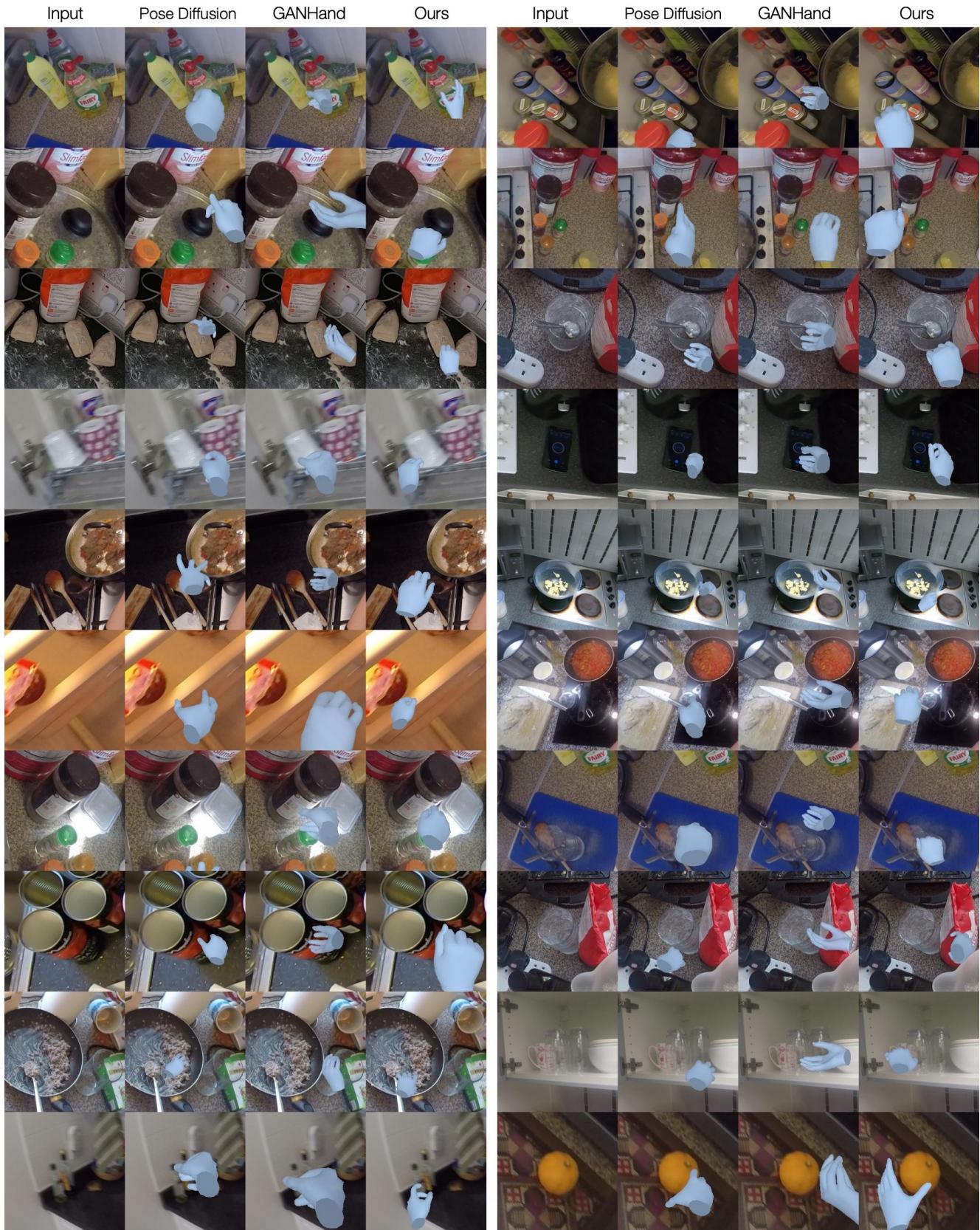
Figure 10. Visualizing the proposed paired data construction applied to EPIC-KITCHEN.

Figure 11. Visualizing more comparisons of the generated HOI images from the proposed method and other image synthesis baselines [36, 41, 68] on the HOI4D dataset.

Figure 12. Visualizing more comparisons of the generated HOI images from the proposed method and other image synthesis baselines [36, 41, 68] on the EPIC-KITCHEN dataset.

Figure 13. Visualizing more comparisons of the extracted 3D hand pose from the proposed method and other 3D affordance baselines [11, 68] on the HOI4D dataset.

Figure 14. Visualizing more comparisons of the extracted 3D hand pose from the proposed method and other 3D affordance baselines on the EPIC-KITCHEN dataset.

Figure 15. Visualizing the ablation of ContentNet for its LDM-based and GLIDE-based implementations (Sec A.2).
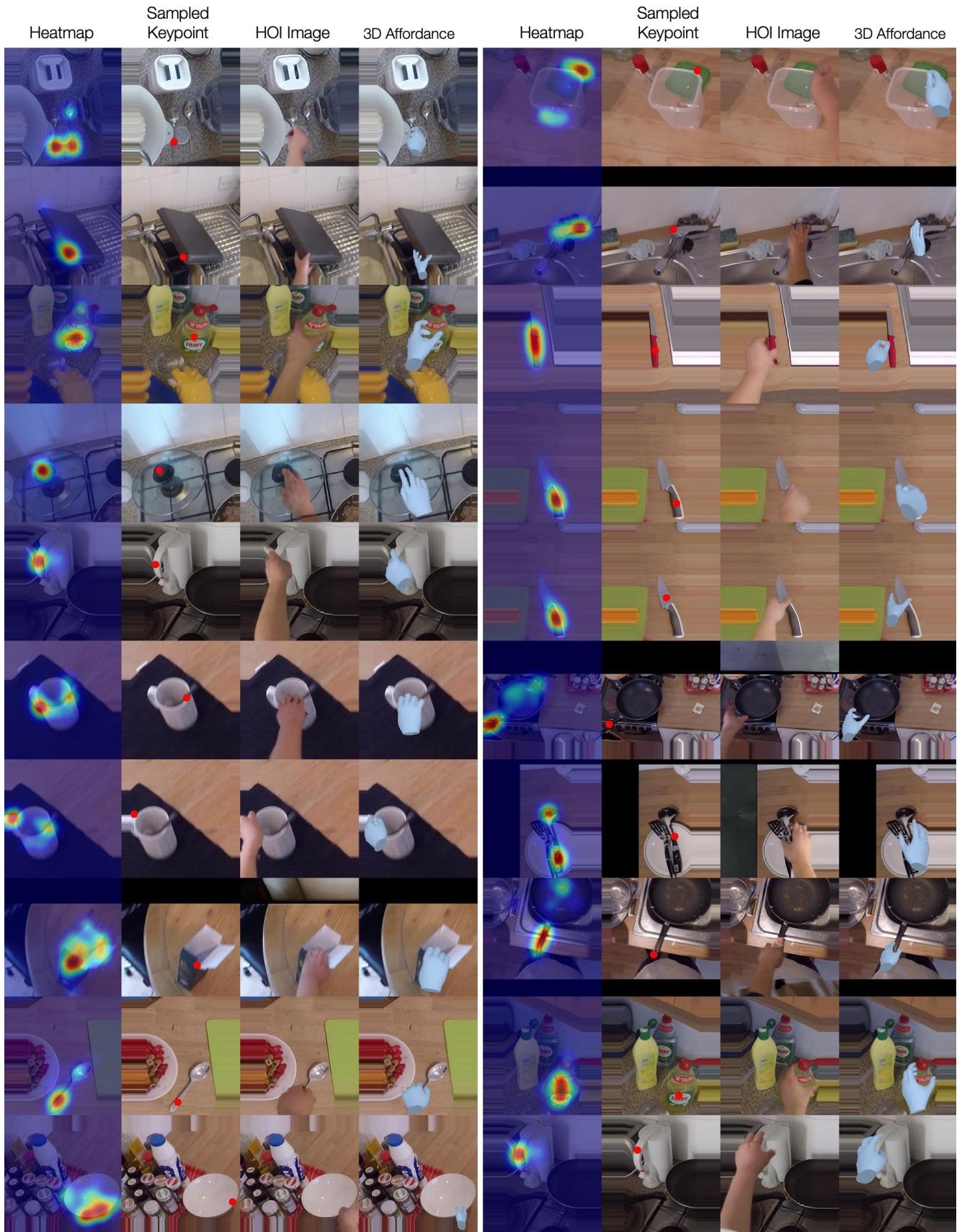
Figure 16. Visualizing more layout editing results.

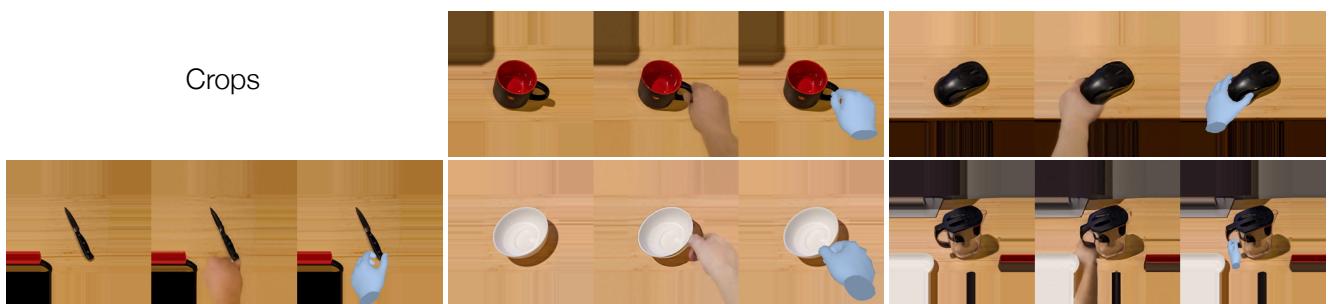Figure 17. Visualizing more results of heatmap-guided synthesis.
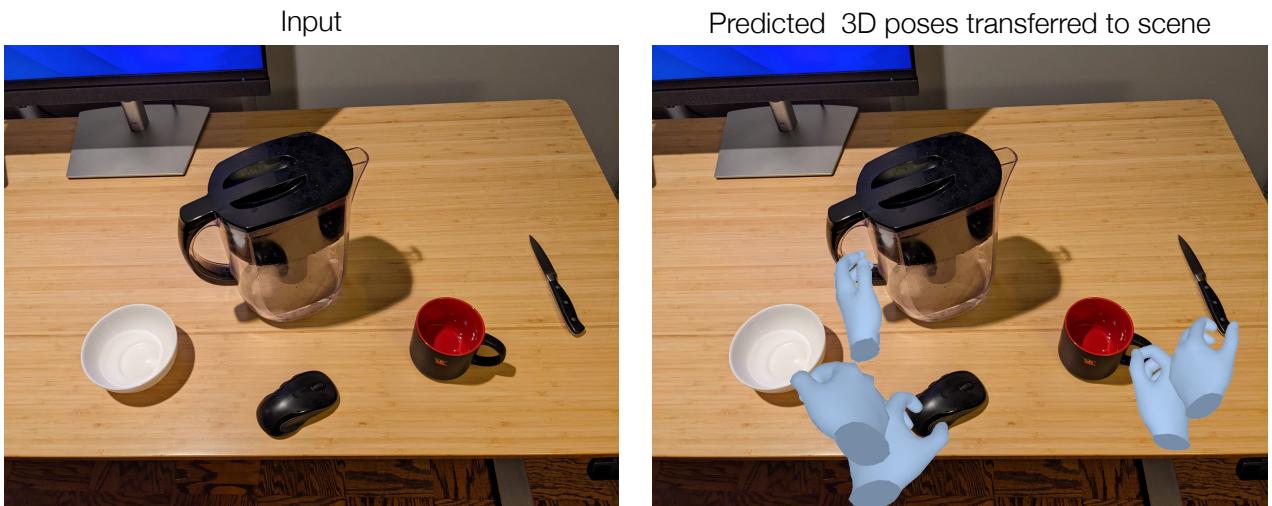
Figure 18. Visualizing more scene integration results with the individual prediction from crops.