

# Diffusion-Guided Reconstruction of Everyday Hand-Object Interaction Clips

Yufei Ye

Poorvi Hebbar

Abhinav Gupta  
Carnegie Mellon University

Shubham Tulsiani

{yufeiy2, phebbar, gabhinav, shubhtuls}@andrew.cmu.edu

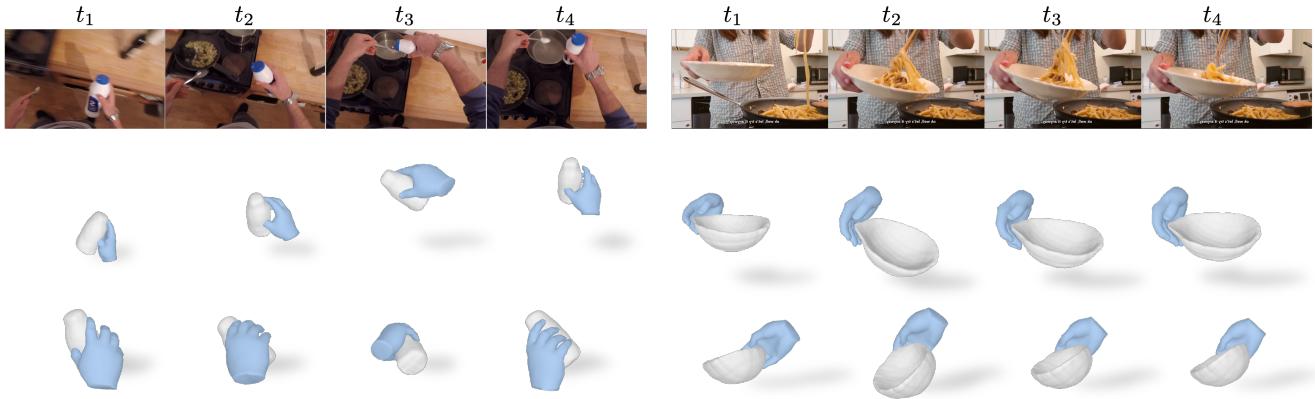
<https://judyye.github.io/diffhoi-www>

Figure 1. Given a video clip depicting a hand-object interaction, we infer the underlying 3D shape of both the hand and the object. **Top:** sampled input frames; **Middle:** reconstruction visualized in the image frame; **Bottom:** reconstruction from a novel view. Please see the website for reconstruction videos.

## Abstract

We tackle the task of reconstructing hand-object interactions from short video clips. Given an input video, our approach casts 3D inference as a per-video optimization and recovers a neural 3D representation of the object shape, as well as the time-varying motion and hand articulation. While the input video naturally provides some multi-view cues to guide 3D inference, these are insufficient on their own due to occlusions and limited viewpoint variations. To obtain accurate 3D, we augment the multi-view signals with generic data-driven priors to guide reconstruction. Specifically, we learn a diffusion network to model the conditional distribution of (geometric) renderings of objects conditioned on hand configuration and category label, and leverage it as a prior to guide the novel-view renderings of the reconstructed scene. We empirically evaluate our approach on egocentric videos across 6 object categories, and observe significant improvements over prior single-view and multi-view methods. Finally, we demonstrate our system’s ability to reconstruct arbitrary clips from YouTube, showing both 1<sup>st</sup> and 3<sup>rd</sup> person interactions.

## 1. Introduction

Our hands allow us to affect the world around us. From pouring the morning coffee to clearing the dinner table, we continually use our hands to interact with surrounding objects. In this work, we pursue the task of understanding such everyday interactions in 3D. Specifically, given a short clip of a human interacting with a rigid object, our approach can infer the shape of the underlying object as well as its (time-varying) relative transformation w.r.t. an articulated hand (see Fig. 1 for sample results).

This task of recovering 3D representations of hand-object interactions (HOI) has received growing interest. While initial approaches [4, 16, 22, 42, 75] framed it as 6-DoF pose task estimation for known 3D objects/templates, subsequent methods have tackled the reconstruction of apriori unknown objects [24, 33, 89]. Although single-view 3D reconstruction approaches can leverage data-driven techniques to reconstruct HOI images [24, 89, 96], these approaches cannot obtain precise reconstructions given the fundamentally limited nature of the single-view input. On the other hand, current video-based HOI reconstruction methods primarily exploit multi-view cues and rely on purely geometry-driven optimization for reconstruction. As

a result, these methods are suited for in-hand scanning where a user carefully presents exhaustive views of the object of interest, but they are not applicable to our setting as aspects of the object may typically be unobserved.

Towards enabling accurate reconstruction given short everyday interaction clips, our approach (DiffHOI) unifies the data-driven and the geometry-driven techniques. Akin to the prior video-based reconstruction methods, we frame the reconstruction task as that of optimizing a video-specific temporal scene representation. However, instead of purely relying on geometric reprojection errors, we also incorporate data-driven priors to guide the optimization. In particular, we learn a 2D diffusion network which models the distribution over plausible (geometric) object renderings conditioned on estimated hand configurations. Inspired by recent applications in text-based 3D generation [39, 61], we use this diffusion model as a generic data-driven regularizer for the video-specific 3D optimization.

We empirically evaluate our system across several first-person hand-object interaction clips from the HOI4D dataset [44], and show that it significantly improves over both prior single-view and multi-view methods. To demonstrate its applicability in more general settings, we also show qualitative results on arbitrary interaction clips from YouTube, including both first-person and third-person clips.

## 2. Related Works

**Reconstructing Hand-Object Interactions.** Hands and objects inherently undergo mutual occlusions which makes 3D reconstruction extremely ill-posed during interactions. Hence, many works [1, 5, 14, 22, 23, 54, 59, 75, 77, 92] reduce the problem to 6D pose estimation by assuming access to instance-specific templates. Their frameworks of 6D pose optimization can be applied to both videos and images. Meanwhile, template-free methods follow two paradigms for videos and images. On one hand, video-based methods takes in synchronized RGB(D) videos [18, 31, 70, 71, 91] or monocular videos [19, 29, 83] and fuse observation to a canonical 3D representation [51]. This paradigm does not use any prior knowledge and requires all regions being observed in some frames, which is often not true in everyday video clips. On the other hand, methods that reconstruct HOI from single images [7, 8, 24, 33, 89] leverage learning-based prior to reconstruct more general objects. While they are able to generate reasonable per-frame predictions, it is not trivial to aggregate information from multiple views in one sequence and generate a time-consistent 3D shape. Our work is template-free and unifies both geometry-driven and data-driven methods.

**Generating Hand-Object Interactions.** Besides reconstructing the ongoing HOIs, many works have explored generating plausible HOIs in different formulations. Some

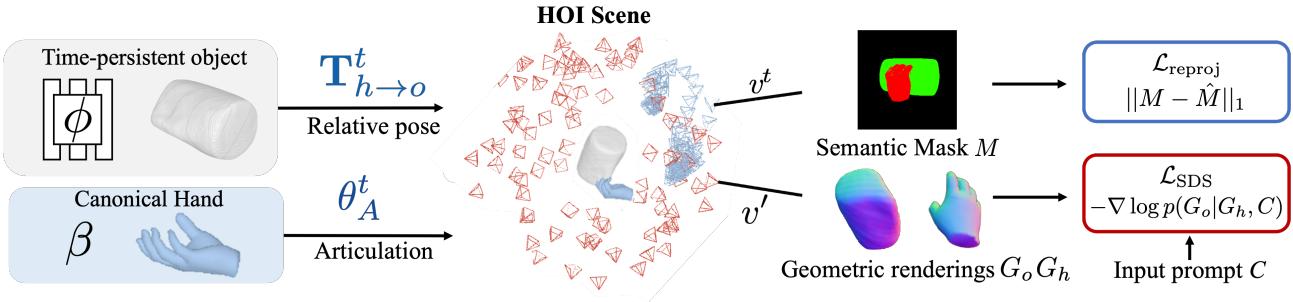
works model their joint distributions [6, 28, 33, 72]. Works that are usually called affordance or grasp prediction study the conditional distribution that generates hands/humans given an object/scenes, in 3D representation [2, 11, 30, 37, 73] or 2D images [10, 36, 80, 90]. Recently, some other works explore the reverse problem that generates plausible scenes for a given human/hand pose [3, 53, 60, 86, 88, 94]. In our work, we follow the latter formulation to learn an image-based generative model of hand-held objects given hands, since hand pose estimation [67] is good enough to bootstrap the system.

**Neural Implicit Fields for Dynamic Scenes.** Neural implicit fields [48, 49, 55] are flexible representation that allows capturing diverse shape with various topology and can be optimized with 2D supervision via differentiable rendering [79, 87]. To extend them to dynamic scenes, a line of work optimizes per-scene representation with additional general motion priors [38, 46, 56, 57, 62]. Though these methods can synthesize highly realistic novel views from nearby angles, they struggle to extrapolate viewpoints [15]. Another line of work incorporate category-specific priors [50, 76, 82, 84, 85] to model articulations. They typically work on single deformable objects of the same category such as quadrupeds or drawers. In contrast, we focus on hand-object interactions across six rigid categories where dynamics are mostly due to changing spatial relations and hand articulations.

**Distilling diffusion models.** Diffusion models [25, 64] have made significant strides in text-to-image synthesis. They can also be quickly adapted to take additional inputs or generate outputs in other domains [27, 93]. Recent works have shown that these image-based diffusion models can be distilled [61, 78] into 3D scene representations for generation or reconstruction [13, 32, 41, 47, 61, 81, 95]. Inspired by them, we also adopt a conditional diffusion model for guiding 3D inference of HOIs. However, instead of learning prior over appearance, we use a diffusion model to learn a prior over a-modal geometric renderings.

## 3. Method

Given a monocular video of a hand interacting with a rigid object, we aim to reconstruct the underlying hand-object interaction, *i.e.*, the 3D shape of the object, its pose in every frame, along with per-frame hand meshes and camera poses. We frame the inference as per-video optimization of an underlying 3D representations. While the multiple frames allow leveraging multi-view cues, they are not sufficient as the object of interests is often partially visible in everyday video clips, due to limited viewpoints and mutual occlusion. Our key insight is to incorporate both view consistency across multiple frames and a data-driven prior



**Figure 2. Method Overview:** We decompose the HOI scene into 1) a rigid time-persistent implicit field  $\phi$  for the object, 2) hand meshes parameterized by hand shape  $\beta$  and articulations  $\theta_A^t$ , and 3) their time-varying relative poses  $T_{h \rightarrow o}^t$ . We define the camera poses  $T_{c \rightarrow h}^t$  in the hand frame. The scene representation is optimized with respect to a reprojection term from the original views  $v^t$  and a data-prior term from novel views  $v'$ .

of the HOIs geometry. The learned interaction prior captures both category priors, *e.g.* mugs are generally cylindrical, and hand priors, *e.g.* pinched fingers are likely to hold thin handles. We train a conditional diffusion model for the prior that guides the HOI to be reconstructed during per-video optimization.

More specifically, given a monocular video  $\hat{I}^t$  with corresponding hand and object masks  $\hat{M}^t \equiv (\hat{M}_h^t, \hat{M}_o^t)$ , we aim to optimize a HOI representation (Sec. 3.1) that consists of a time-persistent implicit field  $\phi$  for the rigid object, a time-varying morphable mesh for the hand  $H^t$ , the relative transformation between hand and object  $T_{h \rightarrow o}^t$ , and time-varying camera poses  $T_{c \rightarrow h}^t$ . The optimization objective consists of two terms (Sec. 3.3): a reprojection error from the estimated original viewpoint and data-driven prior term that encourages the object geometry to appear more plausible given category and hand information when looking from another viewpoint. The prior is implemented as a diffusion model conditioned on a text prompt  $C$  about the category and renderings of the hand  $\pi(H)$  with geometry cues (Sec. 3.2). It denoises the rendering of the object  $\pi(O)$  and backpropagates the gradient to the 3D HOI representation by score distillation sampling (SDS) [61].

### 3.1. HOI Scene Representation

**Implicit field for the object.** The rigid object is represented by a time-persistent implicit field  $\phi$  that can handle unknown topology and has shown promising results when optimizing for challenging shapes [79, 85, 87]. For every point in the object frame, we use multi-layer perceptrons to predict the signed distance function (SDF) to the object surface,  $s = \phi(\mathbf{X})$ .

**Time-varying hand meshes.** We use a pre-defined parametric mesh model MANO [66] to represent hands across frames. The mesh can be animated by low-dimensional parameters and thus can better capture more structured mo-

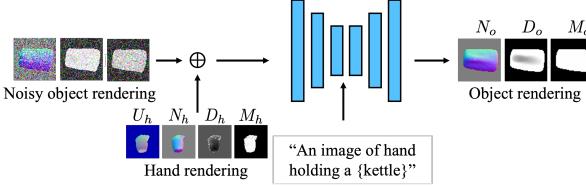
tions, *i.e.* hand articulation. We obtain hand meshes  $H^t$  in a canonical hand wrist frame by rigging MANO with a 45-dim pose parameters  $\theta_A^t$  and 10-dim shape parameters  $\beta$ , *i.e.*  $H^t = \text{MANO}(\theta_A^t, \beta)$ . The canonical wrist frame is invariant to wrist orientation and only captures finger articulations.

**Composing to a scene.** Given the time-persistent object representation  $\phi$  and a time-varying hand mesh  $H^t$ , we then compose them into a scene at time  $t$  such that they can be reprojected back to the image space from the cameras. Prior works [21, 23, 59] typically track 6D object pose directly in the camera frame  $T_{c \rightarrow o}$  which requires an object template to define the object pose. In our case, since we do not have access to object templates, the object pose in the camera frame is hard to estimate directly. Instead, we track object pose with respect to hand wrist  $T_{h \rightarrow o}^t$  and initialize them to identity. It is based on the observation that the object of interest usually moves together with the hand and undergoes “common fate” [69]. A point in the rigid object frame can be related to the predicted camera frame by composing the two transformations, camera-to-hand  $T_{c \rightarrow h}^t$  and hand-to-object  $T_{h \rightarrow o}^t$ . For notation convention, we denote the implicit field transformed to the hand frame at time  $t$  as  $\phi^t(\cdot) \equiv \phi(T_{h \rightarrow o}^t(\cdot))$ . Besides modeling camera extrinsics, we also optimize for per-frame camera intrinsics  $\mathbf{K}^t$  to account for zoom-in effect, cropping operation, and inaccurate intrinsic estimation.

In summary, given a monocular video with corresponding masks, the parameters to be optimized are

$$\phi, \beta, \theta_A^t, T_{h \rightarrow o}^t, T_{c \rightarrow h}^t, \mathbf{K}^t \quad (1)$$

**Differentiable Rendering.** To render the HOI scene into an image, we separately render the object (using volumetric rendering [87]) and the hand (using mesh render-



**Figure 3. Geometry-informed Diffusion Model:** The diffusion model takes in a noisy geometry rendering of the object, the geometry rendering of the hand, and a text prompt, to output the denoised geometry rendering of objects.

ing [43, 58]) to obtain geometry cues. We then blend their renderings into HOI images by their rendered depth.

Given an arbitrary viewpoint  $v$ , both differentiable renders can render geometry images including mask, depth, and normal images, *i.e.*  $G_h \equiv (M_h, D_h, N_h)$ ,  $G_o \equiv (M_o, D_o, N_o)$ . To compose them into a semantic mask  $M_{HOI}$  that is later used to calculate the reprojection loss, we softly blend the individual masks by their predicted depth. Similar to blending two-layer surfaces of mesh rendering, the final semantic masks can be computed by alpha blending:  $M = B(M_h, M_o, D_h, D_o)$ . Please refer to supplementary material for the full derivation of the blending function  $B$ .

### 3.2. Data-Driven Prior for Geometry

When observing everyday interactions, we do not directly observe all aspects of the object because of occlusions and limited viewpoint variability. Despite this, we aim to reconstruct the 3D shape of the full object. To do so, we rely on a data-driven prior that captures the likelihood of a common object geometry given its category and the hand interacting with it  $p(\phi^t | H^t, C)$ . More specifically, we use a diffusion model which learns a data-driven distribution over geometry rendering of objects given that of hands and category.

$$\log p(\phi^t | H^t, C) \approx \mathbb{E}_{v \sim V} \log p(\pi(\phi^t; v) | \pi(H^t; v), C) \quad (2)$$

where  $v \sim V$  is a viewpoint drawn from a prior distribution,  $C$  as category label and  $\pi$  as rendering function. Since this learned prior only operates in geometry domain, there is no domain gap to transfer the prior across daily videos with complicated appearances. We first pretrain this diffusion model with large-scale ground truth HOIs and then use the learned prior to guide per-sequence optimization (Sec. 3.3).

**Learning prior over a-modal HOI geometry.** Diffusion models are a class of probabilistic generative models that gradually transform a noise from a tractable distribution (Gaussian) to a complex (*e.g.* real image) data distribution. Diffusion models are supervised to capture the likelihood by de-noising corrupted images. During training, they take

in corrupted images with a certain amount of noise  $\sigma_i$  along with conditions and learn to reconstruct the signals [26]:

$$\mathcal{L}_{DDPM}[\mathbf{x}; \mathbf{c}] = \mathbb{E}_{\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), i} \|\mathbf{x} - D_\psi(\mathbf{x}_i, \sigma_i, \mathbf{c})\|_2^2 \quad (3)$$

where  $\mathbf{x}_i$  is a linear combination of signal  $\mathbf{x}$  and noise  $\epsilon$  while  $D_\psi$  is the denoiser.

In our case, as shown in Fig. 3, the diffusion model denoises the a-modal geometry rendering of an object given text prompt and hand. Additionally, the diffusion model is also conditioned on the rendering of uv-coordinate of MANO hand  $U_h$  because it can better disambiguate if the hand palm faces front or back. More specifically, the training objective is  $\mathcal{L}_{\text{diff}} = \mathcal{L}_{DDPM}[G_o; C, G_h, U_h]$ . The text prompt comes from a text template: “an image of a hand holding {category}”.

**Implementation Details.** When we train the diffusion model with the rendering of ground truth HOI, we draw viewpoints with rotation from the uniform distribution in  $SO(3)$ . We use the backbone of a text-to-image model [52] with cross attention and modify it to diffuse 5-channel geometry images (3 for normal, 1 for mask and 1 for depth). We initialize the weights from the image-conditioned diffusion model [52] pretrained with large-scale text-image pairs. The additional channels in the first layer are loaded from the average of the pretrained weights.

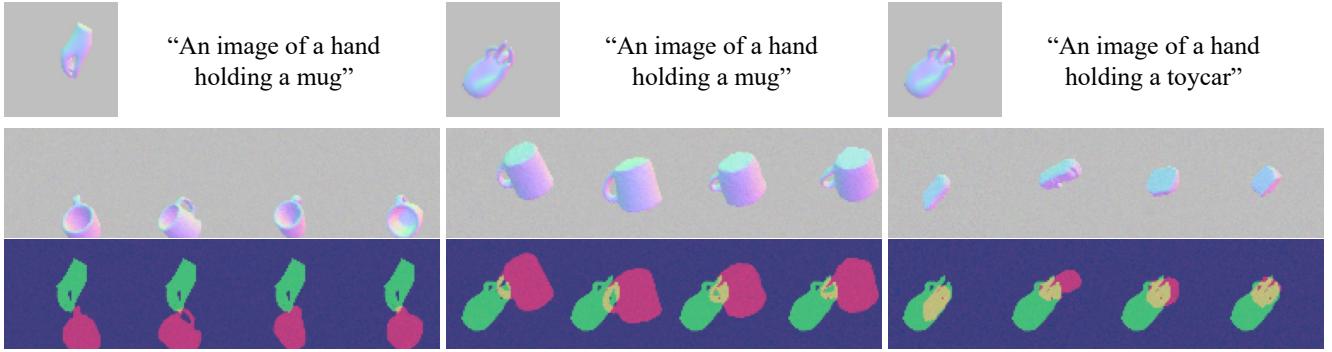
### 3.3. Reconstructing Interaction Clips in 3D

After learning the above interactions prior, at inference time when given a short monocular clip with semantic masks of hand and object, we optimize a per-sequence HOI representation to recover the underlying hand-object interactions. We do so by differentiable rendering of the 3D scene representation from the original views and from random novel views. The optimization objectives consist of the following terms.

**Reprojection error.** First, the HOI representation is optimized to explain the input video. We render the semantic mask of the scene from the estimated cameras for each frame and compare the rendering of the semantic masks (considering hand-object occlusion) with the ground truth masks:  $\mathcal{L}_{\text{reproj}} = \sum_t \|M^t - \hat{M}^t\|_1$

**Learned prior guidance.** In the meantime, the scene is guided by the learned interaction prior to appear more likely from a novel viewpoint following Scored Distillation Sampling (SDS) [61]. SDS treats the output of a diffusion model as a critic to approximate the gradient step towards more likely images without back-propagating through the diffusion model for compute efficiency:

$$\mathcal{L}_{SDS} = \mathbb{E}_{v, \epsilon, i} [w_i \|\pi(\phi^t) - \hat{G}_o^i\|_2^2] \quad (4)$$



**Figure 4. Generations from conditional diffusion model:** Given the geometry rendering of hand  $G_h$  (only showing surface normals) and a text prompt  $C$ , we visualize 4 different generations from the diffusion model. Middle row shows the generated surface normal of the objects and bottom row visualizes the generated object masks overlaid on the given hand masks. Note the left and middle column share the same text condition while middle and right column share the same hand condition.

where  $\hat{G}_o^i$  is the reconstructed signal from the pre-trained diffusion model. Please refer to relevant works [47, 61] or supplementary for full details.

**Other regularization.** We also include two regularization terms: one Eikonal loss [17] that encourages the implicit field  $\phi$  to be a valid distance function  $\mathcal{L}_{\text{eik}} = \|\nabla_X \phi^2 - 1\|^2$ , and another temporal loss that encourages the hand to move smoothly with respect to the object  $\mathcal{L}_{\text{smooth}} = \sum_t \|T_{h \rightarrow o}^t H^t - T_{h \rightarrow o}^{t-1} H^{t-1}\|_2^2$

**Initialization and training details.** While the camera and object poses are learned jointly with object shape, it is crucial to initialize them to a coarse position [40]. We use FrankMocap [67], an off-the-shelf hand reconstruction system, to initialize the hand parameters, camera-to-hand transformations, and camera intrinsic. More specifically, FrankMocap predicts finger articulation  $\theta_A^t$ , wrist orientation  $\theta_w^t$ , and a weak perspective camera. The last two are used to compute camera-to-hand transformation and intrinsics of a full perspective camera. See appendix for derivation. We initialize the object implicit field to a coarse sphere [87] and the object poses  $T_{h \rightarrow o}^t$  to identity such that the initial object is roughly round hand palm.

The per-frame hand pose estimation sometimes fails miserably in some challenging frames due to occlusion and motion blur. We run a lightweight trajectory optimization on wrist orientation to correct the catastrophic failure. The optimization objective encourages smooth joint motion across frames while penalizing the difference to the per-frame prediction, *i.e.*  $\mathcal{L} = \|H(\mathbf{x}^t) - H(\hat{\mathbf{x}}^t)\| + \lambda \|H(\mathbf{x}^{t+1}) - H(\mathbf{x}^t)\|$  where  $\lambda$  is 0.01. Please see appendix for full details.

## 4. Experiment

We first train the diffusion model on the egocentric HOI4D [44] dataset and visualize its generations in Sec-

tion 4.1. Then, we evaluate the reconstruction of hand-object interactions quantitatively and qualitatively on the held-out sequences and compare DiffHOI with two model-free baselines (Section 4.2). We then analyze the effects of both category-prior and hand-prior respectively, ablate the contribution from each geometry modality, and analyze its robustness to initial prediction errors (Section 4.3). In Section 4.4, we discuss how DiffHOI compares with other template-based methods. Lastly, in Section 4.5, we show that our method is able to reconstruct HOI from in-the-wild video clips both in first-person and from third-person view.

**Dataset and Setup.** HOI4D is an egocentric dataset consisting of short video clips of hand interacting with objects. It is collected under controlled environments and recorded by head-wear RGBD cameras. Ground truth is provided by fitting 6D pose of scanned objects to the RGBD videos. We use all of the 6 rigid object categories in portable size (mug, bottle, kettle, knife, toy car, bowl). To train the diffusion model, we render one random novel viewpoint for each frame resulting in 35k training points. We test the object reconstruction on held-out instances, two sequences per category. All of baselines and our method use the segmentation masks from ground truth annotations and the hand poses from the off-the-shelf prediction system [67] if required.

For in-the-wild dataset, we test on clips from EPIC-KITCHENS [12] videos and casual YouTube videos downloaded from the Internet. The segmentation masks are obtained using an off-the-shelf video object segmentation system [9].

### 4.1. Visualizing Data-Driven Priors

We show conditional generations by the pre-trained diffusion model in Fig. 11. Given the geometry rendering of hand (only visualizing surface normals), as well as a text prompt, we visualize 4 different generations from the diffusion model. Middle row shows the generated surface normal of the object and bottom row visualizes the generated

Table 1. **Comparison with baselines:** We compare our method along with prior works HHOR [29] and iHOI [89] on the HOI4D dataset and report object reconstruction error in  $F@5\text{mm}$  and  $F@10\text{mm}$  scores and Chamfer Distance ( $CD$ ).

	Mug			Bottle			Kettle			Bowl			Knife			ToyCar			Mean		
	$F@5$	$F@10$	$CD$	$F@5$	$F@10$	$CD$	$F@5$	$F@10$	$CD$	$F@5$	$F@10$	$CD$									
HHOR [29]	0.18	0.37	7.0	0.26	0.56	3.1	0.12	0.30	11.3	0.31	0.54	4.2	0.71	0.93	0.6	0.26	0.59	1.9	0.31	0.55	4.7
iHOI [89]	0.44	0.71	2.1	0.48	0.77	1.5	0.21	0.45	6.3	0.38	0.64	3.1	0.33	0.68	2.8	0.66	0.95	0.5	0.42	0.70	2.7
Ours (DiffHOI)	<b>0.64</b>	<b>0.86</b>	<b>1.0</b>	<b>0.54</b>	<b>0.92</b>	<b>0.7</b>	<b>0.43</b>	<b>0.77</b>	<b>1.5</b>	<b>0.79</b>	<b>0.98</b>	<b>0.4</b>	0.50	<b>0.95</b>	0.8	<b>0.83</b>	<b>0.99</b>	<b>0.3</b>	<b>0.62</b>	<b>0.91</b>	<b>0.8</b>

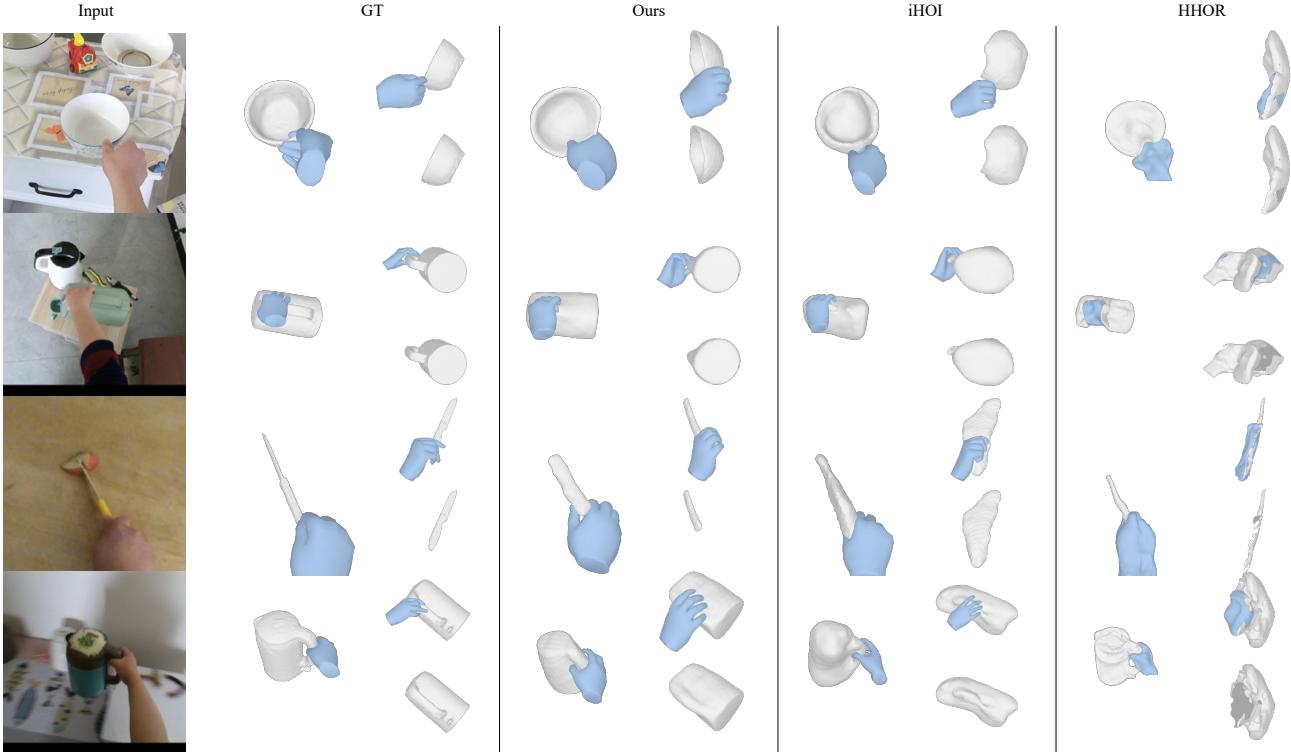


Figure 5. **Qualitative evaluation on HOI4D:** We show reconstruction by our method (DiffHOI) along with two baselines [29, 89] in the image frame (left) and another novel view with (top right) or without (bottom right) hand. Please see project website for reconstruction videos.

object masks overlayed on top of the given hand mask, for a better view of the hand-object relations. Note that left and middle column condition on the same text prompts while middle and right column conditions on the same hand pose. Please see appendix for additional examples and visualizations of all modalities.

The generated object match the category information in the prompt while the generations are diverse in position, orientation, and size. Yet, all of the hand-object interactions are plausible, *e.g.* different generated handles all appear at the tip of the hand. Comparing middle and right examples, different category prompts lead to different generations given the same hand rendering. With the same prompt but different hands (left and middle), the generated objects flip the orientation accordingly. In summary, Fig. 11

indicates that the learned prior is aware of both the hand prior and the category-level prior hence being informative to guide the 3D reconstruction from clips.

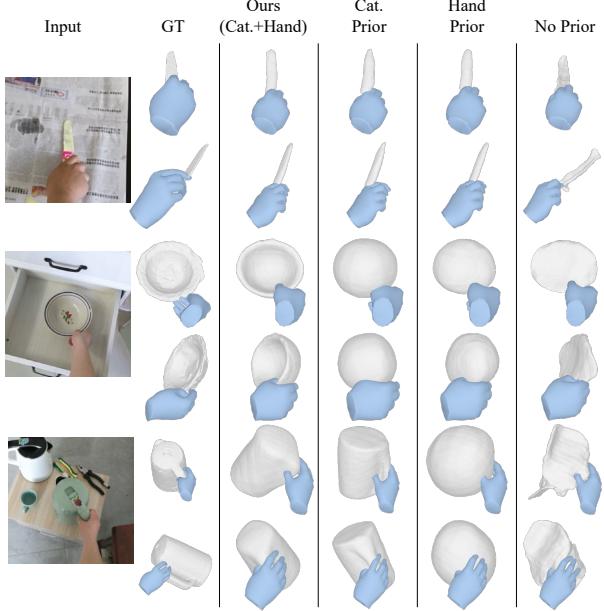
## 4.2. Comparing Reconstructions of HOI4D

**Evaluation Metric.** We evaluate the object reconstruction errors. Following prior works [20, 29], we first align the reconstructed object shape with the ground truth by Iterative Closest Point (ICP), allowing scaling. Then we compute Chamfer distance ( $CD$ ), F-score [74] at  $5\text{mm}$  and  $10\text{mm}$  and report mean over 2 sequences for each category. Chamfer distance focuses on the global shapes more and is affected by outliers while F-score focuses on local shape details at different thresholds [74].

**Baselines.** While few prior works tackle our challenging setting – 3D HOI reconstruction from casual monocular

**Table 2. Analysis of the effect of data-driven priors:** Quantitative results on HOI4D for object reconstruction error in the object-centric frame ( $F@5$ ,  $F@10$ ,  $CD$ ) and for hand-object alignment in the hand frame ( $CD_h$ ). We compare our method with ablations that does not use prior, or use other variants of diffusion models that only conditions on hand or category.

	$F@5$	$F@10$	$CD$	$CD_h$
No prior	0.47	0.73	2.7	<b>37.0</b>
Hand prior	0.39	0.65	2.8	55.0
Category prior	0.56	0.87	1.6	85.2
Ours	<b>0.62</b>	<b>0.91</b>	<b>0.8</b>	48.7



**Figure 6. Ablation Study:** We show reconstruction in the image frame (top) and from a novel view (bottom) by our method along with ablations using other variants of diffusion models that only conditions on category or hand, and one that does not use prior.

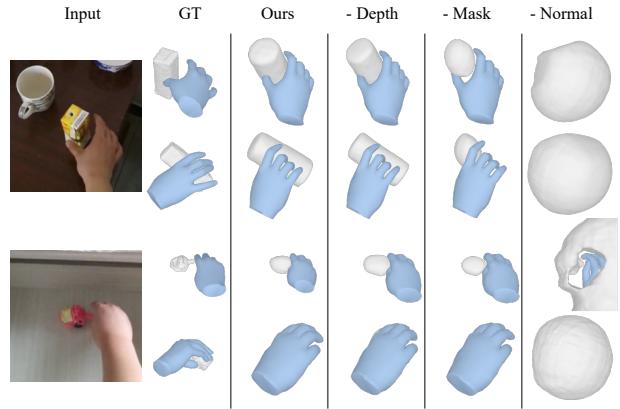
clips without knowing the templates, the closest works are two template-free methods from Huang *et al.* [29] (HHOR) and Ye *et al.* [89] (iHOI).

HHOR is proposed for in-hand scanning. It optimizes a deformable semantic implicit field to jointly model hand and object. HHOR captures the dynamics by a per-frame warping field while no prior is used during optimization. iHOI is a feed-forward method and reconstructs 3D objects from single-view images by learning the hand prior between hand poses and object shapes. The method does not leverage category-level prior and do not consider time-consistency of shapes. We finetune their pretrained model to take in segmentation masks. We evaluate their result by aligning their predictions with ground truth for each frame and report the average number across all frames.

**Results.** We visualize the reconstructed HOI and object shapes from the image frame and a novel viewpoint in

**Table 3. Ablation without surface normal, mask and depth in distillation:** Quantitative results on HOI4D for object reconstruction error in the object-centric frame ( $F@5$ ,  $F@10$ ,  $CD$ ) and for hand-object alignment in the hand frame ( $CD_h$ ). We compare our method with other ablations that do not distill normals, masks, and depths respectively.

	$F@5$	$F@10$	$CD$	$CD_h$
- normal	0.36	0.58	4.3	220.2
- mask	0.56	0.82	1.3	128.1
- depth	<b>0.66</b>	0.90	0.9	88.0
Ours	0.62	<b>0.91</b>	<b>0.8</b>	<b>48.7</b>



**Figure 7. Ablation Study:** We show reconstruction in the image frame (top) and from a novel view (bottom) by our method along with other variants that do not distill on depth, mask, and normals.

Fig. 5. HHOR generates good-looking results from the original view but actually degenerates to a flat surface since it does not incorporate any prior knowledge besides the visual observation. It also cannot decompose the hand and the object on the unobserved side of the scene because HHOR distinguishes them by per-point classification predicted from the neural field, which does not get gradient from the observations. iHOI reconstructs better object shapes and interactions but it is not very accurate as it cannot aggregate information across different frames. Its prediction is not time consistent either (better visualized as videos). In contrast, we are able to reconstruct time-persistent object shapes with time changing hand poses. The reconstructed object is more accurate, *e.g.* knife blade is thinner and the kettle body is more cylindrical.

This is consistent with quantitative results in Tab. 1. HHOR generally performs unfavorably except for knife category. While iHOI performs better, its quality is limited by only relying on information from a single frame. DiffHOI outperforms the baseline methods by large margins in most sequences and performs the best on all three metrics for mean values.

Table 4. Error analysis against hand pose noise: \* marks our unablated method. Numbers in parentheses are per-frame prediction errors before optimization.

	Object Reconstruction			Hand Estimation	
	F@5 ↑	F@10 ↑	CD ↓	MPJPE ↓	AUC ↑
GT	0.68	0.91	0.75	—	—
Prediction*	0.62	0.91	0.77	26.9(28.4)	0.49(0.47)
Pred. Error ×2	0.63	0.87	1.01	40.7(44.6)	0.31(0.27)

### 4.3. Ablation Studies

We ablate our system carefully to analyze the contribution of each component. Besides the object reconstruction errors in the aligned object-centric frame, we further evaluate the hand-object *arrangement* by reporting the Chamfer distance of objects in hand frame, *i.e.*  $CD_h \equiv CD(T_{o \rightarrow h}^t O, \hat{T}_{o \rightarrow h}^t \hat{O})$ . We only report mean value in the main paper. Please refer to supplementary for category-wise results.

**How does each learned prior help?** We analyze how the category and hand priors affect reconstruction by training two more diffusion models conditioned only on text-prompt or hand renderings respectively. We also compare with the variant without optimizing  $\mathcal{L}_{SDS}$  (no prior). As reported quantitatively, we find that *category prior helps object reconstructions while hand prior helps hand-object relation* (Tab. 2). And combining them both results in best performance.

We highlight an interesting qualitative result of reconstructing the bowl in Fig. 6. Neither prior can reconstruct the concave shape on its own – the hand pose alone is not predictive enough of the object shape while only knowing the object to be a bowl cannot make the SDS converge to a consensus direction that the bowl faces. Only knowing *both* can the concave shapes be recovered. This example further highlights the importance of both priors.

### Which geometry modality matters more for distillation?

Next, we investigate how much each geometry modality (mask, normal, depth) contributes when distilling them into 3D shapes. Given the same pretrained diffusion model, we disable one of the three input modalities in optimization by setting its weight on  $\mathcal{L}_{SDS}$  to 0.

As visualized in Fig. 7, the surface normal is the most important modality. Interestingly, the model collapses if not distilling surface normals and even performs worse than the no-prior variant. Without distillation on masks, the object shape becomes less accurate probably because binary masks predict more discriminative signals on shapes. Relative depth does not help much with global object shape but it helps in aligning detailed local geometry ( $F@5$ ) and aligning the object to hand ( $F@10$ ).

**How robust is the system to hand pose prediction errors?** We report the object reconstruction performance

Table 5. Comparison with template-based baseline: Quantitative results on the HOI4D dataset for object reconstruction error in the object-centric frame ( $F@5$ ,  $F@10$ ,  $CD$ ) and for hand-object alignment ( $CD_h$ ). We compare our method with HOMAN [23] with the ground truth template (-GT), with random templates from the training split (and reporting the average), and with furthest template from the ground truth (-furthest).

	F@5 ↑	F@10 ↑	CD ↓		CD <sub>h</sub> ↓
HOMAN-GT	1.00	1.00	0.00		84.3
HOMAN-average	0.76	0.94	0.48		120.9
HOMAN-furthest	0.49	0.78	1.33		157.9
Ours(DiffHOI)	0.62	0.91	0.78		48.7

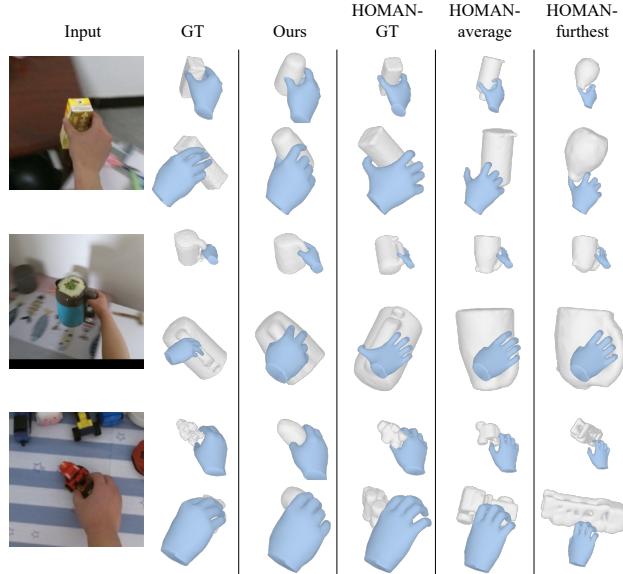


Figure 8. **Comparing with template-based method:** We show reconstruction in the image frame (top) and from a novel view (bottom) by our method and HOMAN [23] when provided with ground-truth templates, a random template, and the most dissimilar template in the training split.

when using GT vs predicted hand pose in Tab. 4, and find that our system is robust to some prediction error. Moreover, even if we artificially degrade the prediction by doubling the error, our performance remains better than the baselines (Tab. 1). We also report the hand pose estimation metrics and find that our optimization improves the initial predictions (in parentheses).

### 4.4. Comparing with Template-Based Methods

We compare with HOMAN [23], a representative template-based method that optimizes object 6D poses and hand articulations with respect to reprojection error and multiple interaction objectives including contact, intersection, distance, relative depth, temporal smoothness, *etc.*

We show quantitative and qualitative results in Tab. 5 and 8. Note that evaluating HOMAN in terms of object

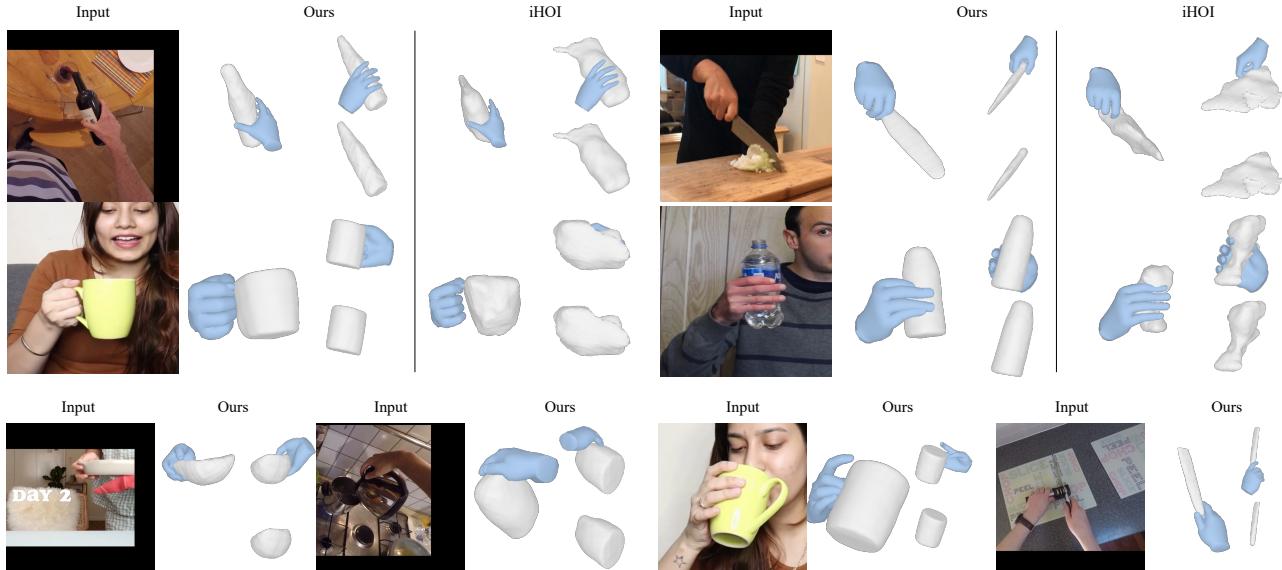


Figure 9. **Qualitative evaluation on in-the-wild video clips:** We show reconstruction by our method (DiffHOI) and iHOI [89] in the image frame (left) and a novel view with (top right) or without (bottom right) hand. Please see project website for reconstruction videos.

reconstruction is equivalent to evaluating templates since the objects are aligned in the object-centric frame. We first report the average object reconstruction errors when optimizing with different templates from training sets. While the gap indicates potential room to improve object shapes for template-free methods, DiffHOI is favorable over some templates in the training set. Nevertheless, when evaluating the objects in the hand frame, DiffHOI outperforms HOMAN by a large margin. The numbers along with visualizations in Fig. 8 indicate that template-based methods, even when optimizes with multiple objectives to encourage interactions, still struggle to place objects in the context of hands, especially for subtle parts like handles. Furthermore, optimizing with random templates degrades  $CD_h$  significantly, highlighting the inherent drawbacks of template-based methods to demand the accurate templates.

#### 4.5. Reconstructing In-the-Wild Video Clips

Lastly, we show that our method can be directly applied to more challenging video clips. In Fig. 9 top, we compare between our method and iHOI [89]. iHOI predicts reasonable shapes from the front view but fails on transparent objects like the plastic bottle since it is never trained on such appearance. In contrast, we transfer better to in-the-wild sequences as the learned prior only take on geometry cues. In Fig. 9 bottom, we visualize more results from our method. By incorporating learned priors, our method is robust to mask prediction inaccuracy, occlusion from irrelevant objects (the onion occludes knife blade), truncation of the HOI scene (bowl at the bottom left), etc. Our method can also work across ego-centric and third-person

views since the learned prior is trained with uniformly sampled viewpoints. The reconstructed shapes vary from thin objects like knives to larger objects like kettles.

## 5. Conclusion

In this work, we propose a method to reconstruct hand-object interactions without any object templates from daily video clips. Our method is the first to tackle this challenging setting. We represent the HOI scene by a model-free implicit field for the object and a model-based mesh for the hand. The scene is optimized with respect to re-projection error and a data-driven geometry prior that captures the object shape given category information and hand poses. Both of these modules are shown as critical for successful reconstruction. Despite the encouraging results, there are several limitations: the current method can only handle small hand-object motions in short video clips up to a few ( $\sim 5$ ) seconds; the reconstructed objects still miss details of shape. Despite the challenges, we believe that our work takes an encouraging step towards a holistic understanding of human-object interactions in everyday videos.

**Acknowledgements.** The authors would thank Di Huang for HHOR comparison. We thank Dandan Shan, Sudeep Dasari for helping with EPIC-KITCHENS datasets. We also thank Sudeep Dasari, Hanzhe Hu, Helen Jiang for detailed feedback on the manuscript. Yufei was partially supported by the NVIDIA fellowship.

## References

- ## References

[1] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *CVPR*, 2022. 2

[2] 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

[3] Tim Brooks and Alexei A Efros. Hallucinating pose-compatible scenes. In *ECCV*, 2022. 2

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. *ICCV*, 2021. 1

[5] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. *ICCV*, 2021. 2

[6] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, 2021. 2

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

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

[9] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *NeurIPS*, 2021. 5, 14

[10] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *CVPR*, 2018. 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

[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. 5

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

[14] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023. 2

[15] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *NeurIPS*, 2022. 2

[16] 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. 1

[17] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *PMLR*, 2020. 5

[18] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escalano, Rohit Pandey, Jason Dougarian, et al. The re-lightables: Volumetric performance capture of humans with realistic relighting. *ToG*, (6), 2019. 2

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

[20] Shreyas Hampali, Tomás Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-hand 3d object scanning from an rgb sequence. *CVPR*, 2023. 6

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

[22] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020. 1, 2

[23] Yana Hasson, Güл Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *3DV*, 2021. 2, 3, 8, 13, 15

[24] Yana Hasson, Güл Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1, 2

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

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 4

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2021. 2

[28] Hezhen Hu, Weilun Wang, Wengang Zhou, and Houqiang Li. Hand-object interaction image generation. *NeurIPS*, 2022. 2

[29] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia*, 2022. 2, 6, 7, 14

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

[31] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralho-fusion: Neural volumetric rendering under human-object interactions. In *CVPR*, 2022. 2

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

[33] 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. 1, 2

- [34] Diederik P Kingma, J Adam Ba, and J Adam. A method for stochastic optimization. *ICLR*, 2020. 14
- [35] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020. 14
- [36] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *CVPR*, 2023. 2
- [37] Xuetong 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
- [38] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *CVPR*, 2023. 2
- [39] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *CVPR*, 2023. 2
- [40] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, 2021. 5
- [41] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *CVPR*, 2023. 2
- [42] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, 2021. 1
- [43] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, 2019. 4, 13
- [44] 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
- [45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2017. 13
- [46] Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman, William T Freeman, and Michael Rubinstein. Omnimatte: Associating objects and their effects in video. In *CVPR*, 2021. 2
- [47] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360° reconstruction of any object from a single image. In *CVPR*, 2023. 2, 5
- [48] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2
- [49] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *ECCV*, 2020. 2
- [50] Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-sdf: Learning disentangled signed distance functions for articulated shape representation. In *ICCV*, 2021. 2
- [51] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 2
- [52] 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*, 2021. 4, 13
- [53] Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Pose2room: understanding 3d scenes from human activities. In *ECCV*, 2022. 2
- [54] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*, 2011. 2
- [55] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2
- [56] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 2
- [57] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ToG*, 2021. 2
- [58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 4, 13
- [59] Austin Patel, Andrew Wang, Ilija Radosavovic, and Jitendra Malik. Learning to imitate object interactions from internet videos. *arXiv*, 2022. 2, 3
- [60] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *CVPR*, 2023. 2
- [61] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR*, 2022. 2, 3, 4, 5, 14
- [62] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021. 2
- [63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*, 2021. 13
- [64] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2021. 2
- [65] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 13

- [66] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 2017. 3
- [67] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV Workshops*, 2021. 2, 5, 14
- [68] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 14
- [69] Dandan Shan, Richard Higgins, and David Fouhey. Cohesiv: Contrastive object and hand embedding segmentation in video. *NeurIPS*, 2021. 3
- [70] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingyi Yu, and Jingya Wang. Neural free-viewpoint performance rendering under complex human-object interactions. In *MM*, 2021. 2
- [71] Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. Neuralhumanfvv: Real-time neural volumetric human performance rendering using rgb cameras. In *CVPR*, 2021. 2
- [72] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *CVPR*, 2022. 2
- [73] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 2
- [74] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhiwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019. 6
- [75] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, 2019. 1, 2
- [76] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. Implicit mesh reconstruction from unannotated image collections. *arXiv*, 2020. 2
- [77] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 2016. 2
- [78] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023. 2
- [79] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 2, 3
- [80] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 2
- [81] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *ICLR*, 2022. 2
- [82] Fangyin Wei, Rohan Chabra, Lingni Ma, Christoph Lassner, Michael Zollhöfer, Szymon Rusinkiewicz, Chris Sweeney, Richard Newcombe, and Mira Slavcheva. Self-supervised neural articulated shape and appearance models. In *CVPR*, 2022. 2
- [83] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *CVPR*, 2023. 2
- [84] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. *CVPR*, 2023. 2
- [85] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 2, 3
- [86] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*, 2022. 2
- [87] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 2021. 2, 3, 5, 13
- [88] Sifan Ye, Yixing Wang, Jiaman Li, Dennis Park, C Karen Liu, Huazhe Xu, and Jiajun Wu. Scene synthesis from human motion. In *SIGGRAPH Asia*, 2022. 2
- [89] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022. 1, 2, 6, 7, 9
- [90] Yufei Ye, Xuetong Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 2
- [91] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023. 2
- [92] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020. 2, 13, 15
- [93] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv*, 2023. 2
- [94] Qian Zheng, Weikai Wu, Hanting Pan, Niloy Mitra, Daniel Cohen-Or, and Hui Huang. Inferring object properties from human interaction and transferring them to new motions. *Computational Visual Media*, 2021. 2
- [95] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 2
- [96] Berk Çalli, Arjun Singh, Aaron Walsman, Siddhartha S. Srinivasa, P. Abbeel, and Aaron M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. *ICRA*, 2015. 1

# Diffusion-Guided Reconstruction of Everyday Hand-Object Interaction Clips

## Supplementary Materials

Yufei Ye

Poorvi Hebbar

Abhinav Gupta

Shubham Tulsiani

Carnegie Mellon University

In the supplementary materials, we provide more implementation details and experimental results. We discuss the details of differentiable rendering of the HOI scene representation (Sec. A.1), network architectures (Sec. A.2), scored distillation sampling of the pretrained diffusion model (Sec. A.3), and initialization details (Sec. A.5). We also describe how to get 2D segmentation masks from in-the-wild clips (Sec. A.4). Then, we show generation by the diffusion model (Sec. B.1), full quantitative results reported in the main paper (Sec. B.2). Furthermore, we also show supporting evidence that optimizing per-frame object poses (Sec. B.3) and soft blending (Sec. B.4) are both important for better performance. Lastly, we discuss our failure cases in Sec. B.5.

## A. Implementation Details

### A.1. Differentiable Rendering (Sec. 3.1)

Given an HOI scene representation at a certain time  $t$  consisting of an implicit field for the object and a mesh for the hand, we use differentiable volumetric renderer [87] and mesh renderer [43, 58] to get their masks ( $M_o, M_h$ ) and depth ( $D_o, D_h$ ). In order to supervise them with reprojection loss with respect to the ground truth semantic masks, we blend hand and object masks by their predicted depths to obtain the rendered semantic masks  $M \equiv B(M_h, M_o, D_h, D_o)$ .

The soft blending is computed as expected light transported to the cameras, similar to blending two-layer surfaces of in mesh rendering [58]. More specifically, denote  $m_h, d_h, m_o, d_o$  as the value at pixel  $(i, j)$ , e.g.  $m_h \equiv M_h[i, j]$ . For any pixel  $(i, j)$ , the blended value is computed as

$$m = B(m_h, m_o, d_h, d_o) = \frac{\sum_{k=0,1} w_k l_k}{\sum_{k=0,1} w_k + w_{bg}} \quad (5)$$

where subscript  $k$  denotes the sorted value of hand and object according to the predicted depth;  $l_k$  is the one-hot semantic label (all 0 for background).  $w_k$  is the weight com-

puted from depth:

$$w_k = m_k \exp \frac{z_k - \max_{k,i,j} Z_k[i, j]}{\gamma}, z_k = m_k \frac{d^{\text{far}} - d_k}{d^{\text{far}} - d^{\text{near}}} \quad (6)$$

We show in Sec. B.4 that soft blending (with loss in semantic masks) is important for better results and performs favorably to the alternative (hard blending with ordinal depth loss [23, 92]).

### A.2. Network Architectures and Training Details (Sec. 3.1 3.2)

**Implicit field.** We use Multi-Layer Perceptron (MLPs) to implement the neural implicit surface of the object  $\phi$ . We borrow the architecture in the original VolSDF [87] and reduce the network capacity to half as we find it to suffice. More specifically, we stack four-layer blocks of which each is a linear layer with channel dim 64 followed by a Soft-Plus activation. We apply positional encoding to the queried point  $X$  with 6 frequencies.

**Conditional diffusion models.** The backbone of the conditional diffusion model is based on the architecture of the text-to-image inpainting model [52]. More specifically, it is a 16-layer UNet with cross attentions and skip layers. The text condition along with the diffusion step embedding is passed to the bottleneck of the UNet and is fused with the image feature by cross-attention. The text prompt is encoded as CLIP tokens [63].

**Details of training diffusion model.** We train the diffusion model with batch size 8, learning rate  $1e-4$ . We use AdamW [45] optimizer with weight decay 0.01 and train for  $500k$  iterations. We use linear noise schedule [65].

**Details of optimizing HOI scene.** We follow the training setup in a reimplementation<sup>1</sup> of the original paper [87]. We optimize the scene with 1024 rays per step, and set initial

<sup>1</sup><https://github.com/ventusff/neurecon>

Table 6. Full ablation results of object reconstruction: Quantitative results for object reconstruction error using F1@5mm and F1@10mm scores and Chamfer Distance (mm). We compare our method with variants that do not optimize per-frame object poses (Sec.B.3), blend hand and object masks in a hard way (Sec.B.4), or do not distill certain geometry modality (Sec. 4.2, Tab. 4)

	Mug			Bottle			Kettle			Bowl			Knife			ToyCar			Mean		
	F@5	F@10	CD	F@5	F@10	CD	F@5	F@10	CD	F@5	F@10	CD	F@5	F@10	CD	F@5	F@10	CD	F@5	F@10	CD
no prior	0.46	0.73	1.8	0.39	0.65	2.2	0.18	0.39	9.1	0.45	0.73	1.9	0.70	0.93	0.5	0.63	0.92	0.6	0.47	0.73	2.7
hand prior	0.48	0.77	1.4	0.37	0.66	1.6	0.30	0.60	3.4	0.38	0.63	4.2	0.09	0.24	5.8	0.70	0.97	0.4	0.39	0.65	2.8
cat. prior	0.62	0.85	1.1	0.56	0.95	0.6	0.63	0.94	0.7	0.35	0.58	5.8	0.44	0.94	0.8	0.77	0.98	0.4	0.56	0.87	1.6
wo learning pose	0.67	0.86	1.0	0.39	0.85	1.1	0.26	0.62	2.4	0.79	0.99	0.3	0.58	0.95	0.7	0.82	0.99	0.3	0.59	0.88	1.0
hard blending	0.54	0.80	1.4	0.51	0.90	0.8	0.29	0.66	2.5	0.60	0.90	0.8	0.65	0.95	0.6	0.83	0.99	0.3	0.57	0.87	1.1
– mask	0.46	0.74	1.7	0.23	0.51	2.6	0.38	0.72	2.2	0.71	0.96	0.5	0.83	0.98	0.3	0.77	0.99	0.3	0.56	0.82	1.3
– normal	0.48	0.77	1.4	0.21	0.44	3.7	0.25	0.49	5.2	0.38	0.63	3.9	0.10	0.22	11.4	0.75	0.95	0.5	0.36	0.58	4.3
– depth	0.69	0.93	0.6	0.73	0.91	0.8	0.51	0.86	1.2	0.38	0.70	2.1	0.79	0.98	0.4	0.82	0.98	0.3	0.65	0.89	0.9
Ours	0.64	0.86	1.0	0.54	0.92	0.7	0.43	0.77	1.5	0.79	0.98	0.4	0.50	0.95	0.8	0.83	0.99	0.3	0.62	0.91	0.8

Table 7. Full ablation results of HOI alignment: Quantitative results for hand-object alignment using Chamfer distance (mm) in hand frame ( $CD_h$ ). We compare our method with variants that do not optimize per-frame object poses (Sec.B.3), blend hand and object masks in a hard way (Sec.B.4), or do not distill certain geometry modality (Sec. 4.2, Tab. 4).

	Mug	Bottle	Kettle	Bowl	Knife	ToyCar	Mean
no prior	36.0	15.4	58.2	75.7	29.5	7.1	37.0
hand prior	34.5	18.3	57.5	87.5	71.7	60.6	55.0
cat. prior	23.2	75.7	54.4	158.6	164.0	34.9	85.2
wo opt. obj pose	21.0	14.1	41.8	167.1	127.1	33.2	67.4
hard blending	26.1	29.9	89.2	205.8	116.1	59.6	87.8
– mask	36.0	28.5	60.7	504.4	97.9	41.3	128.1
– normal	394.9	284.1	107.9	235.5	286.0	296.6	267.5
– depth	14.6	12.7	45.5	270.6	160.6	24.0	88.0
Ours	18.1	15.3	42.2	101.8	91.6	23.3	48.7

learning rate  $5e - 4$  with exponential learning rate scheduler. We use Adam [34] optimizer and optimize for 50k iterations per scene. Within a batch, we bias the sampled pixels from the background, hand, and object region with probability 0.35, 0.35, 0.3 and linearly interpolate the probability to 0.1, 0.1, 0.8 in order to spend more effective computation on the object of interest, same as HHOR [29]. In the first 100 warm-up iterations, we turn off SDS and only optimize for the reprojection loss and other regularization terms. This will make the optimization more stable.

### A.3. Score Distillation Sampling (Sec. 3.3)

With the pretrained diffusion model, we follow DreamFusion [61] to distill the learned prior to the 3D representation. The main idea is to let the diffusion model denoise the corrupted renderings and treats the denoised output as ‘ground truth’. More specifically, at each optimization step, we randomly sampled a viewpoint with random rotation from  $SO(3)$  and random camera distance. Then, we render the geometry renderings  $G_o, G_h$  from the given viewpoint in resolution 64x64. Next, we corrupt the geometry rendering of the object with some noise  $G_o^i = \sqrt{\bar{\alpha}_i}G_o + \sqrt{1 - \bar{\alpha}_i}\epsilon$  ( $\bar{\alpha}$  is the noise scheduling,  $\epsilon$  is a gaussian noise) and pass it through the diffusion model along with the geometry ren-

dering of the hand and text prompt.

$$\hat{G}_o^i = D_\psi(G_o^i | G_h, C) \quad (7)$$

We set the classifier-free guidance scale to 4, which is different from the original paper where a small guidance scale cannot converge. It is probably because 2D observations provide stronger cues than text thus leading to easier convergence.

### A.4. Obtaining hand-object masks for in-the-wild clips.

While we provide ground truth segmentation masks to all methods on HOI4D, we obtain the segmentation masks by off-the-shelf prediction systems [9,35,68] for in-the-wild clips. More specifically, we first use a hand-object interaction detector [68] to detect the location of the hand and the active object in the first frame. Then, given the detected bounding boxes, we use PointRend [35] to get the corresponding masks. Next, we pass the masks of interest in the first frame to a video object segmentation system STCN [9] and obtain the tracked masks in every frame.

To automatically filter out the clips with undesirable segmentation quality, we run the STCN to track forward and backward in time and calculate the Intersection over Union (IoU) between the initial masks and the masks after tracking back. We use clips with IoU higher than 40% for both hand and object masks.

### A.5. Initialization with Off-the-Shelf Predictions (Sec. 3.3.)

We use an off-the-shelf hand reconstruction [67] to estimate initial camera poses  $T_{c \rightarrow h}^t$ , hand shape parameter  $\beta$ , and hand articulation  $\theta_A^t$ . The off-the-shelf system predicts per-frame 10-dim hand shape parameters  $\beta^t$ , 48-dim hand poses  $\theta^t$ , and a weak perspective camera  $s^t, t_x^t, t_y^t$ . We take the average of shape parameters across all frames to initialize the hand shape parameter. Among the 48-dim predicted hand pose, we use the 45-dim finger articulation  $\theta_A^t$

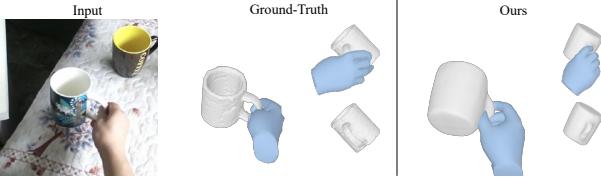


Figure 10. Failure Case

to initialize hand articulation parameter while use the remaining 3-dim wrist orientation  $\theta_w$  as the rotation component of camera pose  $T_{c \rightarrow h}^t$ . The translation component is computed by converting the predicted weak-perspective camera to a full-perspective camera (we use a pinhole camera with a focal length of 1 and the principal point at the center of the frame following Zhang *et al.* [92]). This is to handle large perspective effects, which are common in daily videos of indoor scenes. Given focal length  $f$  and principal points  $p_x, p_y$ , the translation component then becomes  $t^t = ((t_x^t - p_x)/s^t, (t_y^t - p_y)/s^t, f/s^t)$ . To put them together, the initial camera pose in the hand frame is initialized as:

$$T_{c \rightarrow h}^t = [R^t | t^t] = [\text{Rot}(\theta_w^t)] \begin{pmatrix} (t_x^t - p_x)/s^t \\ (t_y^t - p_y)/s^t \\ f/s^t \end{pmatrix] \quad (8)$$

## B. Additional Results

### B.1. Results of diffusion model generation

We show some conditional generations by the pre-trained diffusion model in Fig. 11. Given the geometry rendering of hand (i) of which row 1-4 visualize surface normal, depth, mask, and uv coordinate, as well as a text prompt with category information, we visualize 5 different generations (ii-vi) from the diffusion model. Row 1-3 in col ii-vi shows the generated geometry rendering of the object, and row 4 visualizes overlayed hand and object masks for a better view of the hand-object relations, *i.e.* our model does not output (ii-vi 4). All examples on the left use the ground truth pairs of hand and category information while each example to its right uses another random category but remains hand the same.

As shown in the figure, the generated object matched the category information in the prompt while the generations are diverse in position, orientation, and size. Yet, all of the hand-object interactions are realistic, *e.g.* different generated kettle/mug handles all appear at the tip of the hand. Comparing left and right examples, different category prompts lead to different generations given the same hand rendering. With the same prompt but different hands, the generated objects also change appearance accordingly. For example, in the subfigure [Left A,C], the handles appear at the left when the hand approaches from the left and vice versa.

Fig. 11 indicates that the learned prior is aware of both the hand prior and the category-level prior hence being informative to guide the 3D reconstruction from clips.

### B.2. Category-wise results in ablations (Tab. 4)

In Tab. 4 in the main paper, we only report mean value across all categories due to space limits. We provide quantitative results across all categories in Tab. 6 (object reconstruction) and Tab. 7 (HOI alignment).

### B.3. Ablation: Optimizing vs Fixing Object Pose.

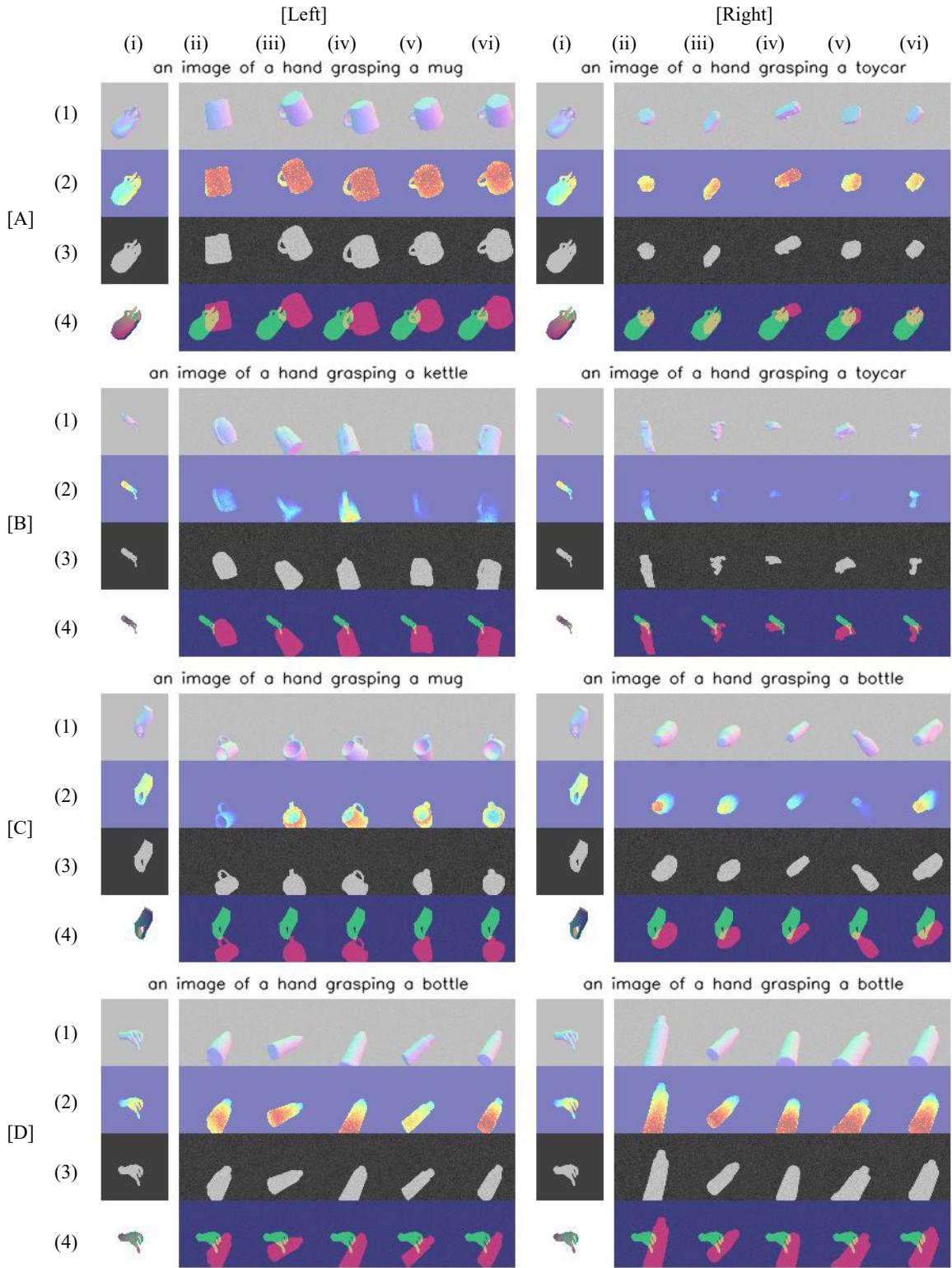
While we observe that the pose of the object in contact relative to hands  $T_{h \rightarrow o}^t$  does not change much, we still optimize per-frame object poses to account for potential relative motion. As reported in Tab. 6, 7 and shown on the project page, allowing changing pose across time improves the performance.

### B.4. Ablation: Soft Blending

Our method obtains the final HOI semantic masks by soft blending hand and object rendering as a weighted sum of the labels where the weight depends on their predicted depth. The alternative way is to select the label of the front surface and apply additional ordinal depth loss. This is common in optimizing the interactions of two template meshes [23, 92]. As shown in the qualitative results on the webpage, the alternative method generates less desirable hand-object relations as the hand intersects with the object. It is consistent with quantitative results in Tab. 6 and 7.

### B.5. Failure Cases

We show one failure case in Fig. 10. The reconstructed mug is in wrong orientation because only semantic masks are used in the reprojection loss. We also struggle with concavity as it is hard to be regularized from only renderings.



**Figure 11. Generations from conditional diffusion model.** Given the geometry rendering of hand (i) (row 1-4 visualizing surface normal, depth, mask, and uv coordinate), as well as a text prompt with category information, we visualize 5 different generations (ii-vi) from the diffusion model. Row 1-3 in col ii-vi shows the generated geometry rendering of the object, and row 4 visualizes overlaid hand and object masks for a better view of the hand-object relations. All examples on the left use the ground truth paired hand and category information while each example to its right uses another random category but remain hand the same.