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 [15] and mesh renderer [6, 9] 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 [9]. 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

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

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 computed from depth:

$$w_k = m_k \exp \frac{z_k - \max_{k,i,j} Z_k[i, j]}{\gamma}, z_k = m_k \frac{d^{\text{far}} - d_k}{d^{\text{far}} - d^{\text{near}}} \quad (2)$$

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 [16, 2]).

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 ϕ. We borrow the architecture in the original VolSDF [15] 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 [8]. 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 [11].

Details of training diffusion model. We train the diffusion model with batch size 8, learning rate $1e-4$. We use AdamW [7] optimizer with weight decay 0.01 and train for $500k$ iterations. We use linear noise schedule [12].

Details of optimizing HOI scene. We follow the training setup in a reimplementation [1] of the original paper [15]. We optimize the scene with 1024 rays per step, and set initial

[1] 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 [4] 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 [3]. 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 [10] 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, ϵ is a gaussian noise) and pass it through the diffusion model along with the geometry rendering of the hand and text prompt.

$$\hat{G}_o^i = D_\psi(G_o^i | G_h, C) \tag{3}$$

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 [14, 1, 5] for in-the-wild clips. More specifically, we first use a hand-object interaction detector [14] to detect the location of the hand and the active object in the first frame. Then, given the detected bounding boxes, we use PointRend [5] to get the corresponding masks. Next, we pass the masks of interest in the first frame to a video object segmentation system STCN [1] 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 [13] to estimate initial camera poses $T_{c\to h}^t$, hand shape parameter β, and hand articulation θ_A^t. The off-the-shelf system predicts

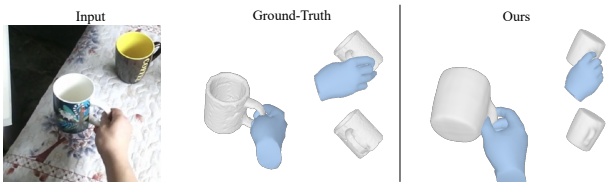

Figure 10: Failure Case

per-frame 10-dim hand shape parameters β^t, 48-dim hand poses θ^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 θ_A^t to initialize hand articulation parameter while use the remaining 3-dim wrist orientation θ_w as the rotation component of camera pose $T_{c\to h}^t$. 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*. [16]). 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 $l^t = ((t_x^t - p_x)/s^t, (t_y^t - p_y)/s^t, f/s^t)$. To put them together, the initial camera pose in the hand frame is initialized as:

$$T_{c\to h}^t = [R^t|l^t] = [\text{Rot}(\theta_w^t)| \begin{pmatrix} (t_x^t - p_x)/s^t \\ (t_y^t - p_y)/s^t \\ f/s^t \end{pmatrix}] \quad (4)$$

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 $T_{h\to o}^t$ 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 [16, 2]. 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.

References

[1] 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.

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

[3] 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.

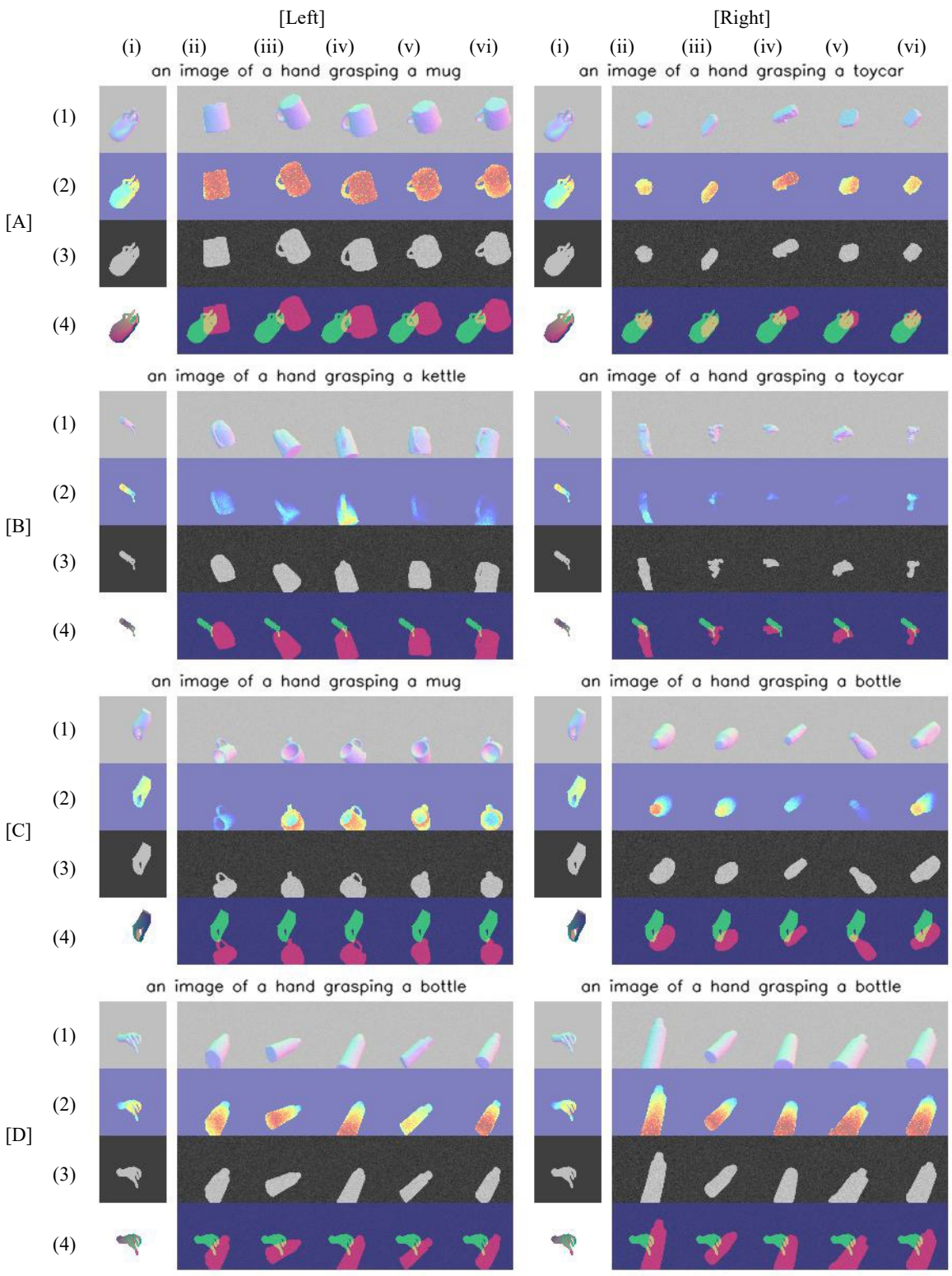

Figure 11: **Generations from conditional diffusion model.** Given the geometry rendering of hand (i) (row 1-4 visualizing surface normal, depth, mask, and uv coordinate), as well as a text prompt with category information, we visualize 5 different generations (ii-vi) from the diffusion model. Row 1-3 in col ii-vi shows the generated geometry rendering of the object, and row 4 visualizes overlayed 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.

[4] Diederik P Kingma, J Adam Ba, and J Adam. A method for stochastic optimization. *ICLR*, 2020.

[5] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020.

[6] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, 2019.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2017.

[8] 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.

[9] 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.

[10] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR*, 2022.

[11] 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.

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

[13] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV Workshops*, 2021.

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

[15] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 2021.

[16] 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.