ForeHOI: Feed-forward 3D Object Reconstruction from Daily Hand-Object Interaction Videos

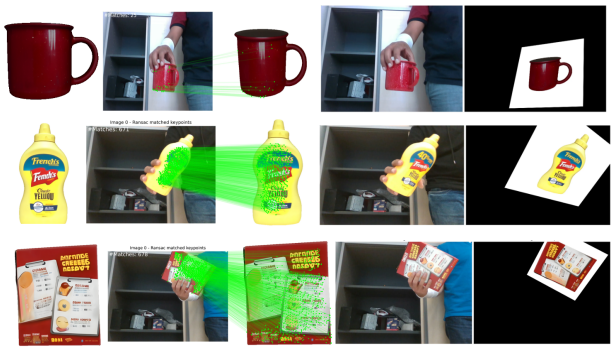
Supplementary Material

In the supplementary material, we provide more details of our model architectures and additional results. We first present details of our object pose estimation module and hand object alignment methods in Section 1. Then in Section 2, we discuss additional experimental results.

1. Methods

1.1. Texture results.

We have trained the texture generation module using our dataset as the second stage of the generation pipeline. Following prior works, we prioritize emphasizing geometric accuracy in the main paper. While our textures are less photo-realistic than those of optimization-based NeRF methods, the preserved color layout is sufficient for reliable 2D correspondence matching. Below are textures for the three cases in main paper.



1.2. Hand object alignment

Since we mainly focus on the object shape and pose estimation, we didn't include the hand object alignment module in the main paper. It's worth noting that in most previous works [1, 6, 11], only hand poses are optimized in the hand-object alignment stage, while object pose and shape are typically frozen. Following MagicHOI [11], we utilize a visible contact alignment strategy to avoid the influence of unreliable object surfaces from heavily occluded areas. Specifically, we first decode our input hand features into 3D hand mesh through mano [8] layers and the optimization process in WiLoR [7] as the initial hand mesh. Then, we utilize a mask projection detection mechanism to mark the visible hand mesh vertices as reliable vertices $\mathcal{V}_h$. For each reliable hand vertex, we use ray tracing to locate the corresponding object contact points $\mathcal{V}_o$. We thus get a set of reliable hand-object pairs $\mathcal{V}_h$, $\mathcal{V}_o$. Finally, the hand translation $t_h \in \mathbb{R}^3$ and scale $s \in \mathbb{R}$ are optimized with the following loss:

$$\mathcal{L}_{contact} = \sum_{i=0}^{M} \left\| \mathcal{V}_{\mathrm{h}}^{i} - \mathcal{V}_{\mathrm{o}}^{i} \right\|$$

$$\mathcal{L}_{\mathrm{kpoints}} = \sum_{i=0}^{M} \left\| \mathbf{P}_{h}^{i} - \mathbf{P}_{o}^{i} \right\|$$

$$\mathcal{L}_{\mathrm{vsmooth}} = \sum_{t=1}^{N} \sum_{i=0}^{M} \left\| \mathcal{V}_{t}^{i} - \mathcal{V}_{t-1}^{i} \right\|_{2}^{2}$$

$$\mathcal{L}_{\mathrm{ho}} = \lambda_{\mathrm{contact}} \mathcal{L}_{\mathrm{contact}} + \lambda_{\mathrm{kpoints}} \mathcal{L}_{\mathrm{kpoints}} + \lambda_{\mathrm{vsmooth}} \mathcal{L}_{\mathrm{vsmooth}} \tag{1}$$

Where M is the number of paired vertices, $\mathcal{V}_t^i$ is the i-th vertex at frame t, $\mathbf{P}$ is the projected points of vertices in 2D space. The hyperparameter $\lambda_{contact}$ is set to 200.0, and $\lambda_{kpoints} = 20.0$, $\lambda_{vsmooth} = 20.0$.

2. Experimental Results

2.1. Comparison with general methods

Since our approach is similar to modern image-to-3D diffusion models, We further qualitatively compare our method with SOTA image-to-3D generative models, including ReconViaGen [2], and Hunyuan3D-3.0 multi-view version [10]. We first mask the object out as RGBD images. Then, for Hunyuan3D, we uniformly sampled 4 images since it can only accept 4 inputs; for ReconViaGen, we feed all images to it. As shown in Fig. 1, although ReconViaGen possesses a strong reconstruction prior and can adapt to image data from arbitrary viewpoints, it lacks the ability to complete occluded areas by other objects, like a hand in the image, while only being capable of reconstructing self-occluded back surfaces. This leads to incomplete reconstruction outcomes. On the other hand, Hunyuan3D-3.0, despite being trained on massive datasets and exhibiting strong object completion capabilities, struggles to interpret input images from arbitrary viewpoints, resulting in artifacts reminiscent of multi-object splicing. Furthermore, both methods exhibit significant distortions in object geometry, a phenomenon similarly observed in our ablation study in the main paper, model trained using TRELLIS's object datasets. This further substantiates a commonly overlooked fact: a considerable bias exists in existing 3D data, wherein the majority of objects are aligned along the gravitational direction, thus leading to 3D generation models trained with these datasets failing to generalize to hand-held

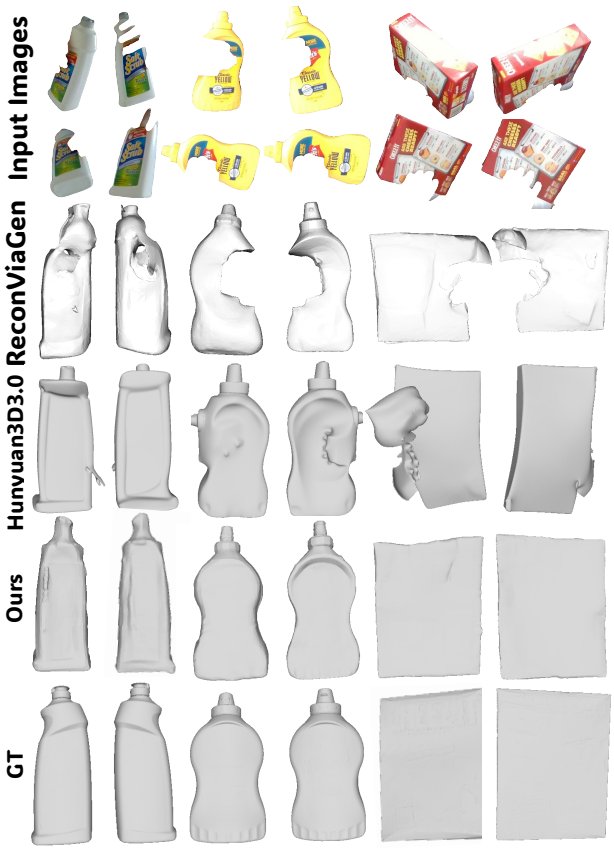


Figure 1. **More Qualitative results comparing with Hunyuan3D-3.0 [10] and ReconViaGen [2] dataset. Zoom in for better visualization in detail.**

objects. This observation underscores the importance of our proposed dataset.

2.2. Hand pose results

Figure 2. **Hand-object aligned visualization on HO3D [3] dataset. Zoom in for better visualization in detail.**

As described in Sec. 1, we align the hand to the object condition on the object shape and pose resulting from ForeHOI. We visualize the hand-object alignment results in Fig. 2. Our hand object alignment result is comparable, even a bit better than MagicHOI [11], noting that MagicHOI utilizes object pose inputs. In the HO3D [3] dataset, the object poses are clipped from the complete pose trajectory obtained via SOTA Structure-from-Motion methods [9] on the full dense sequences. Therefore, these poses can be considered nearly ground-truth.

2.3. Our Datasets

As shown in Fig 3, our dataset contains rich object shapes, poses, orientations, and photorealistic hands compared with previous datasets [4, 5].

Figure 3. **Qualitative comparisons with ObMan [4] and AffordPose [5] dataset.** Noting that ObMan is a single-image dataset, we have much more videos(vids) and objects(objs) compared to the previous largest dataset.

References

[1] Ayce Idil Aytekin, Helge Rhodin, Rishabh Dabral, and Christian Theobalt. Follow my hold: Hand-object interaction reconstruction through geometric guidance, 2025. 1

[2] Jiahao Chang, Chongjie Ye, Yushuang Wu, Yuantao Chen, Yidan Zhang, Zhongjin Luo, Chenghong Li, Yihao Zhi, and Xiaoguang Han. Reconviagen: Towards accurate multi-view 3d object reconstruction via generation, 2025. 1, 2

[3] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 2

[4] Yana Hasson, Gül 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

[5] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14713–14724, 2023. 2

[6] Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7037–7047, 2025. 1

[7] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, 2024. 1

[8] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6), 2017. 1

[9] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2

[10] Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material, 2025. 1, 2

[11] Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. Magichoi: Leveraging 3d priors for accurate hand-object reconstruction from short monocular video clips. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1, 2