OAKINK: A Large-scale Knowledge Repository for Understanding Hand-Object Interaction

[1,2]Lixin Yang*, [1]Kailin Li*, [1]Xinyu Zhan*, [1]Fei Wu, [1]Anran Xu, [1]Liu Liu, [1,2]Cewu Lu†
[1]Shanghai Jiao Tong University, China [2]Shanghai Qi Zhi Institute, China
{siriusyang, kailinli, kelvin34501, legendary, xuanran, liuliu1993, lucewu}@sjtu.edu.cn

Appendices

Contents

A *Oak* base Details;
B Data Annotation Details;
C More Dataset Analysis;
D Implementation: IntGen and HoverGen;
E Perceptual Survey for Generation Tasks;
F Additional Benchmark Results;
F.1 Hand Mesh Recovery: Other Splits;
F.2 Unseen Out-of-domain Object;
F.3 More Visualization;
G Discussion on Personally Identifiable Data;

A. *Oak* Base Details

In this section, we provide the details of the **O**bject **A**ffordance **K**nowledge base (*Oak* base), covering the lists of total 32 categories and 30 *attribute* phrases in Tab. 1.

maniptool	knife, screwdriver, hammer, wrench, toothbrush, pen, frying pan, drill, pincer, scissors, stapler, mug, teapot, cup, can, box, bowl, wineglass, cylinder bottle, trigger sprayer, lotion bottle
functool	eyeglasses, headphones, binoculars, game controller, lightbulb, camera, flashlight, mouse, phone, apple, banana, donut
Attribute phrases	contain sth, cover sth, pump out sth, cut sth, stab sth, flow in/out sth, tighten sth, loosen sth, clamp sth, brush sth, trigger sth, observe sth, point to sth, shear sth, attach to sth, connect sth, knock sth, spray sth, no function; hold by sth, screwed by sth, unscrewed by sth, pressed by sth, handled by sth, plug by sth, unplug by sth, squeeze by sth, pour out by sth

Table 1. The **categories** and **attribute phrases** in our *Oak* base

B. Data Annotation Details

This section is a supplementary of the Sec. 3.2.3: **Hand Pose and Geometry**. Given the manually labeled 2D hand keypoints, we aim to solve the pose $\boldsymbol{\theta} \in \mathbb{R}^{16\times3}$, shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ parameters and the wrist's position $\boldsymbol{P}_{h,0} \in \mathbb{R}^3$ of a 3D hand. These parameters will drive a 3D hand model by a differentiable MANO layer: $\mathcal{M}(\cdot)$ [4]:

$$\boldsymbol{V}_h, \boldsymbol{P}_h = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) + \boldsymbol{P}_{h,0} \tag{1}$$

where $\boldsymbol{P}_h \in \mathbb{R}^{21\times3}$ is the hand joints' 3D position, and $\boldsymbol{V}_h \in \mathbb{R}^{778\times3}$ is the hand mesh vertices' 3D position. The objective cost function for solving $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{P}_{h,0}$ consists of 5 terms.

Reprojection Error. First, we want the 2D projections of the 3D hand joints $\boldsymbol{P}_h$ to match its 2D keypoints annotation $\hat{\boldsymbol{p}}$. Let the subscript j and v be the joint's ID and view's ID, we have the reprojection cost:

$$E_{\text{repj}} = \frac{1}{\sum w_{j,v}} \sum_{v=1}^{4} \sum_{j=1}^{21} w_{j,v} \left\| \boldsymbol{K}_v \boldsymbol{T}_v \boldsymbol{P}_{h,j} - \hat{\boldsymbol{p}}_{j,v} \right\|_2^2 \tag{2}$$

The gradients from E_{repj} will back propagate to $\boldsymbol{P}_h$ and then update the $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{P}_{h,0}$.

Geometry Consistency. Second, we want the 3D geometry model of hand and object to be consistent with their real-world observation: no interpenetration would occur. Hence, we introduce the second cost function: interpenetration loss. We acquire the object's sign distance field: $\mathcal{O}$ from its scanned model, transform the $\mathcal{O}$'s pose from MoCap system to the camera system, and calculate the sign distance value of a 3D hand vertex $\boldsymbol{V}_{h,i}$ to $\mathcal{O}$. The interpen-

etration cost penalizes those hand vertices inside the object surface (with negative sign distance values).

$$E_{\text{intp}} = \sum_{\boldsymbol{V}_{h,i}} -\min\Big(\text{SDF}_{\mathcal{O}}(\boldsymbol{V}_{h,i}), 0\Big), \tag{3}$$

The gradients from E_{intp} will back propagate to each $\boldsymbol{V}_{h,i}$ and then update the $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{P}_{h,0}$.

Silhouette Constraint. Third, we want the contour projection of hand and object models to match the visual cues. Hence, we introduce a binary silhouette cost. We first acquire the hand and object's binary mask ($\mathcal{B}_h$ and $\mathcal{B}_o$) from the recorded images. This process is automatic. We filter out the background pixels through green-screen and depth image matting. The remaining foreground pixels are the union of $\mathcal{B}_h$ and $\mathcal{B}_o$. Then, we render the 3D hand and object mesh on an image as silhouette and penalize the per-pixel misalignment between the rendered silhouette and the binary mask.

$$E_{\text{sh}} = \sum_{\text{all pix.}} \underbrace{f(\boldsymbol{\mathcal{V}}_o \cup \boldsymbol{\mathcal{V}}_h)}_{\text{detached}} \cap BCE\Big\{f(\boldsymbol{\mathcal{V}}_o \cup \boldsymbol{\mathcal{V}}_h), (\mathcal{B}_h \cup \mathcal{B}_o)\Big\} \tag{4}$$

In this equation, the $f(\cdot)$ is a differentiable rendering function [1]; the $(\boldsymbol{\mathcal{V}}_o \cup \boldsymbol{\mathcal{V}}_h)$ is the composited mesh model of hand $\boldsymbol{\mathcal{V}}_h$ and object $\boldsymbol{\mathcal{V}}_o$; the $f(\boldsymbol{\mathcal{V}}_o \cup \boldsymbol{\mathcal{V}}_h)$ is the rendered silhouette image of hand and object model; the $(\mathcal{B}_h \cup \mathcal{B}_o)$ is the union of hand and object binary mask; and the $BCE\{,\}$ is the binary cross entropy loss function; The gradients from E_{sh} will back propagate to $\boldsymbol{\mathcal{V}}_h$ and then update the $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{P}_{h,0}$.

Anatomical Constraint. Forth, we want the MANO hand pose to satisfy the anatomical constraints of human hand. Hence, we borrow the axial adaptations from Yang *et al*. [7] and constrain the rotation axes and angles.

$$E_{\text{anat}} = \sum_{j \in \text{all}} \Big(\boldsymbol{a}_j \cdot \mathbf{n}_j^t + \max\big((\phi_j - \frac{\pi}{2}), 0\big)\Big) + \sum_{j \notin \text{MCP}} \boldsymbol{a}_j \cdot \mathbf{n}_j^s, \tag{5}$$

where the $\boldsymbol{a}_j$ and ϕ_j denote the axial and angular components of the j-th joint's rotation, the $\mathbf{n}_j^t$ and $\mathbf{n}_j^s$ are the predefined *twist* and *splay* direction, and"MCP" indicates the five Metacarpal joints. The gradients from E_{anat} will back propagate to each joint's axis-angle and then update the $\boldsymbol{\theta}$.

Temporal Smoothing. The above cost functions can only improve the per-frame precision of 3D hand annotations. However, frame-by-frame smoothness is also critical to improving our annotation quality. Hence, We want the solved 3D hand poses to be continuous in the time domain. We adopt a low-pass filter (*e.g*. Kalman Filters) to post-process the poses $\boldsymbol{\theta}$ and wrist positions $\boldsymbol{P}_{h,0}$ across the entire image sequence.

C. More Dataset Analysis

Hand Pose Distribution. We project the interacting hand poses into an embedded space yield from t-SNE [6]. The poses that transferred from the same *OakInk*-Core pose are painted in the same color. From the box in Fig. 1, we can see that the similar interacting hand poses with different objects are mapped to adjacent in the embedded 2D space. From the circles in Fig. 1, we can conclude that the different grasping types are away from each other.

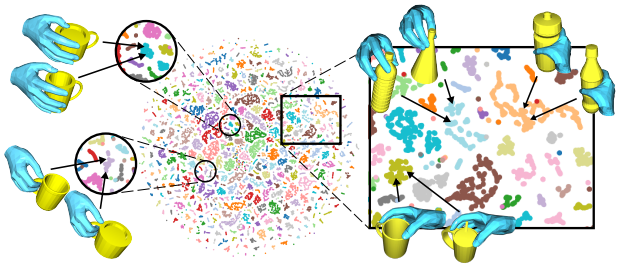

Figure 1. t-SNE embedding of hand poses. We randomly select 20 colors to visualize the clustered poses.

Contact Distribution. We provide the contact heatmaps on example objects that reveal the frequencies of contact among all interactions. Fig. 2 shows such heatmaps on six *Oak* base categories. We see that the "hot" area (red) that denotes the high frequency of contact is consistent with the object affordance we described.

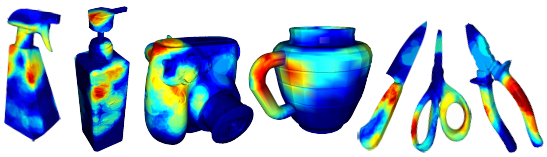

Figure 2. Heatmaps of contact frequency on object surface.

D. Implementation: IntGen and HoverGen

The architecture of IntGen (Fig. 4) and HoverGen (Fig. 5) model are modified from the original GrabNet [5] (Fig. 3) design.

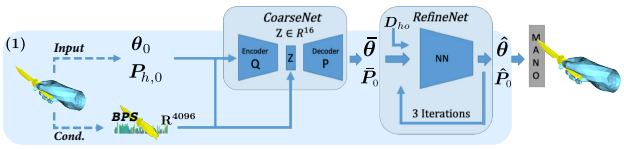

Figure 3. **GrabNet** [5]: the original design.

In the IntGen task, we select three intents: *use*, *hold* and *hand-out*, map the intents' word string to a real-valued word vector, and train the networks with the intent vector as the additional input. During training, poses within different intents will be mapped to different areas in the latent

pose space: $\mathcal{Z} \in \mathbb{R}^{16}$. The training loss functions in IntGen are identical to those in GrabNet, including standard conditional VAE losses (KL divergence and weight regularization), mesh reconstruction losses (hand vertices and mesh edge loss), and physical quality losses (penetration and contact loss). We train the IntGen on category-level data in *OakInk*-Shape, including the *mug*, *camera*, *trigger sprayer* and *lotion bottle*. The training process lasts 1,000 epochs, with the mini batch size 32 and initial learning rate of 1×10^{-3}. The learning rate decays by a factor of 0.5 at every 200 epochs.

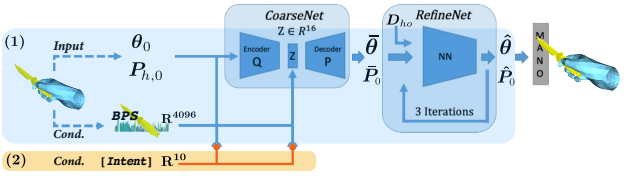

Figure 4. **IntGen**: the intent-based grasp generation network design.

As shown in Fig. 5, in the HoverGen task, we provide the root rotation: $\boldsymbol{\theta}_0^\star$ and root position: $\boldsymbol{P}_{h,0}^\star$ of the giver's hand as the additional inputs for CoarseNet, and the Chamfer distance: $\boldsymbol{D_{hh}} \in \mathbb{R}^{778}$ from the original giver's hand to the predicted receiver's hand as an additional input for RefineNet. As a results, the HoverGen models learns a receiving hand's embedding space, $\mathcal{Z}$, conditioned on the object shape and the giver's hand root pose. At inference time, given an unseen object shape and the giver's hand root pose: $(\boldsymbol{\theta}_0^\star, \boldsymbol{P}_{h,0}^\star)$, we sample a vector from $\mathcal{Z}$ and decode a receiver hand pose to complete a human-to-human handover. The training loss in HoverGen model includes all the losses in IntGen model, plus an L1 loss on the Chamfer distance $\boldsymbol{D_{hh}}$ w.r.t. the ground-truth $\hat{\boldsymbol{D}}_{\boldsymbol{hh}}$, a Chamfer distance from the original giver's hand to the ground-truth receiver's hand. We train the HoverGen model 1,000 epochs with a mini-batch size 256 and initial learning rate of 1×10^{-3}, decaying a half at every 200 epochs.

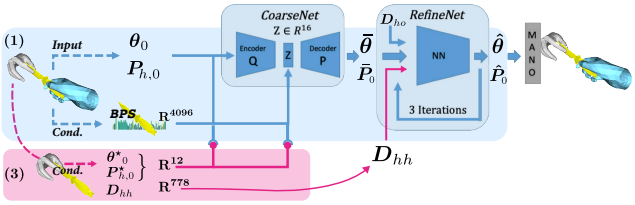

Figure 5. **HoverGen**: the handover generation network design. The giver's hand is paint in **gray** and the receiver's hand in **blue**.

E. Perceptual Survey for Generation Tasks

To investigate the general audience's opinion about the predicted pose of the generation tasks: GrabNet, IntGen, and HoverGen, we conducted three perceptual surveys on the Amazon Mechanical Turk (AMT). In each survey, we show four views of each predicted hand-object interaction and ask the audiences to give their opinion about a statement (*e.g.* "*the hand is interacting naturally with the object*"). The audiences are asked to rate the statement with a 5-level Likert scale ("strongly agree" corresponds to grade 5 and "strongly disagree" corresponds to grade 1). The layout of the perceptual surveys on GrabNet, IntGen, and HoverGen are shown in Fig. 6.

F. Additional Benchmark Results

F.1. Hand Mesh Recovery: Other Splits

Apart from the default split **`SP0`** (split by views) in the main text, we also provide another two data splits and the HMR benchmark results for *OakInk*-Image.

- **`SP1` (subjects split).** (train/val/test: 6/1/5). We split the *OakInk*-Img by subjects. The subjects recorded in the test split will not appear in the train split.
- **`SP2` (objects split).** (train/test: 70%/5%/25%). We split the *OakInk*-Img by objects. The objects that have been grasped in the test split will not appear in the train split.

Splits	Methods	MPJPE↓ (AUC↑)	MPVPE↓
`SP1`	I2L-MeshNet [3]	18.04 (*0.641*)	18.08
	HandTailor [2]	15.72 (*0.792*)	16.31
`SP2`	I2L-MeshNet [3]	15.79 (*0.733*)	15.87
	HandTailor [2]	14.14 (*0.846*)	14.81

Table 2. **HMR results in *mm***. AUC are shown in parentheses.

F.2. Unseen Out-of-domain Object

We refer the objects in *OakInk*-Shape test set as unseen in-domain objects, indicating that it may have similar counterparts included in the training set. In this part, we are also interested in the performance of our generation tasks on the unseen **out-of-domain** objects. We choose the Stanford bunny, a general 3D test model, as an illustrative prototype of **out-of-domain** object. We test the GrabNet and HoverGen model on the Stanford bunny and provide the generated grasps and receiving poses in Fig. 7. Both the GrabNet and HoverGen model are trained on our *OakInk*-Shape training set. The results show that through training on the *OakInk*-Shape, GrabNet and HoverGen can synthesize realistic and prehensile interactions for general objects.

F.3. More Visualization

We provide more qualitative results of the benchmark results of HMR task in Fig. 8, HOPE task in Fig. 9, GraspGen (GrabNet) in Fig. 10: top, IntGen in Fig. 10: middle, and HoverGen in Fig. 10: bottom.

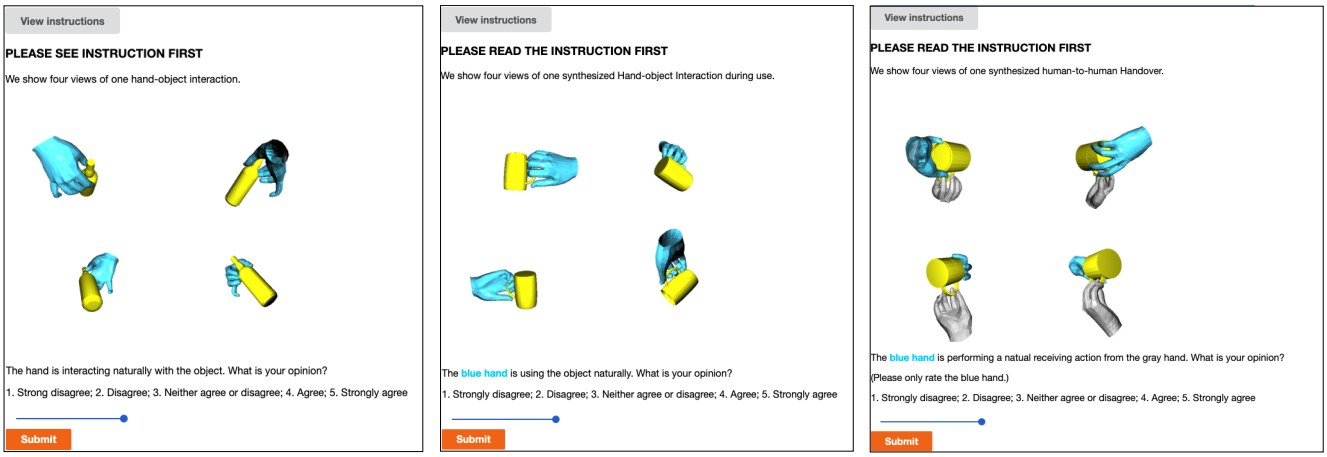

Figure 6. **The layout of three perceptual surveys on AMT**.
Left: GrabNet (statement: *the hand is interacting naturally with the object*);
Middle: IntGen (statement: *the blue hand is using the object naturally*);
Right: HoverGen (statement: *the blue hand is performing a natural receiving action from the gray hand*)

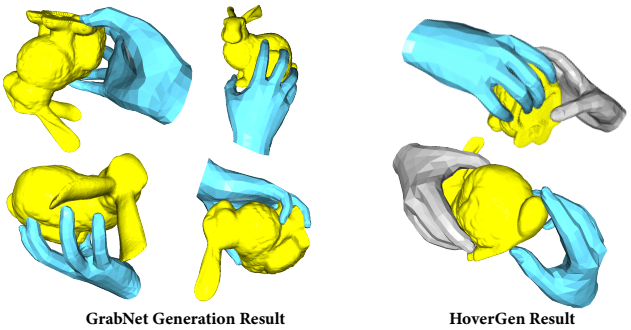

Figure 7. Generation results on **unseen out-of-domain** objects.

G. Discussion on Personally Identifiable Data

We collect hand-object interaction data on 12 human subjects recruited through a third-party crowd-sourcing company. In the collection process, their actions will be recorded in video sequences by the MulCam system. We ensure that the data collecting process meets the ethics requirements through the following announcements:

- The third-party crowd-sourcing company warrants appropriate IRB approval (or equivalent, based on local government requirement) are obtained. The company name and warranties are withheld based on the anonymous submission guidelines.
- All the subjects involved in data collection are required to sign a contract with the third-party crowd-sourcing company, involving permission on the portrait usage, the acknowledgment of data usage, and payment policy. During the data collecting process, all subjects are paid by the hour.
- All the subjects involved in the data collecting process acknowledge that the collected data will only be intended for academic and permitted commercial usages.
- We ensure all the subjects involved in the data collecting process are willing to share the personal-related data, including actions, skin tones, body/hand shapes, *etc*.
- We require all the subjects not to dress in revealing or offensive clothes during the data collection process.
- Upon the release of the dataset, we will desensitize all samples in the dataset by blurring the subjects' faces (if any), tattoos, rings, or any other accessories that may reveal the subjects' identity.

References

[1] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018. 2

[2] Jun Lv, Wenqiang Xu, Lixin Yang, Sucheng Qian, Chongzhao Mao, and Cewu Lu. HandTailor: Towards high-precision monocular 3d hand recovery. In *BMVC*, 2021. 3

[3] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 3

[4] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 2017. 1

[5] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 2

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008. 2

[7] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *ICCV*, 2021. 2

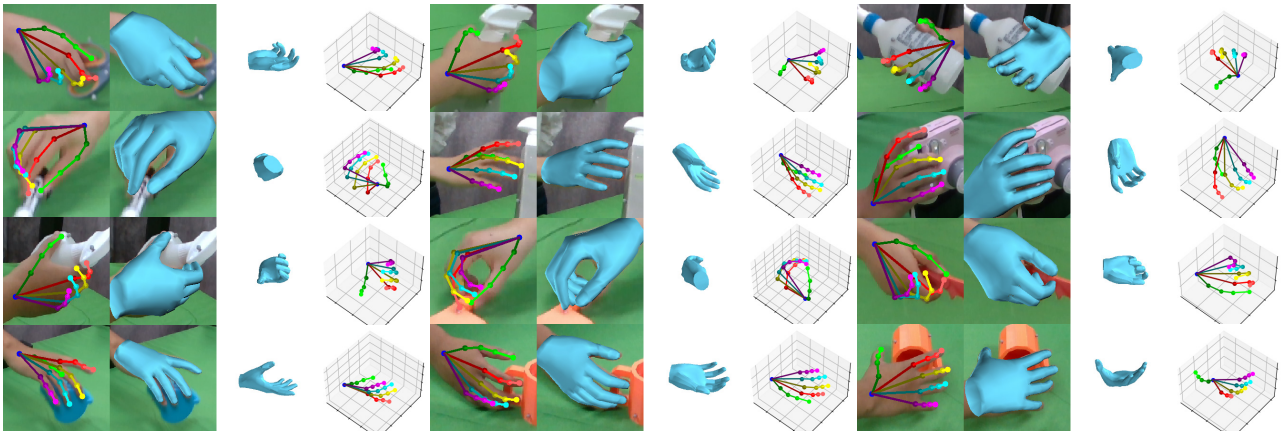

Figure 8. **More qualitative results** on HMR task.

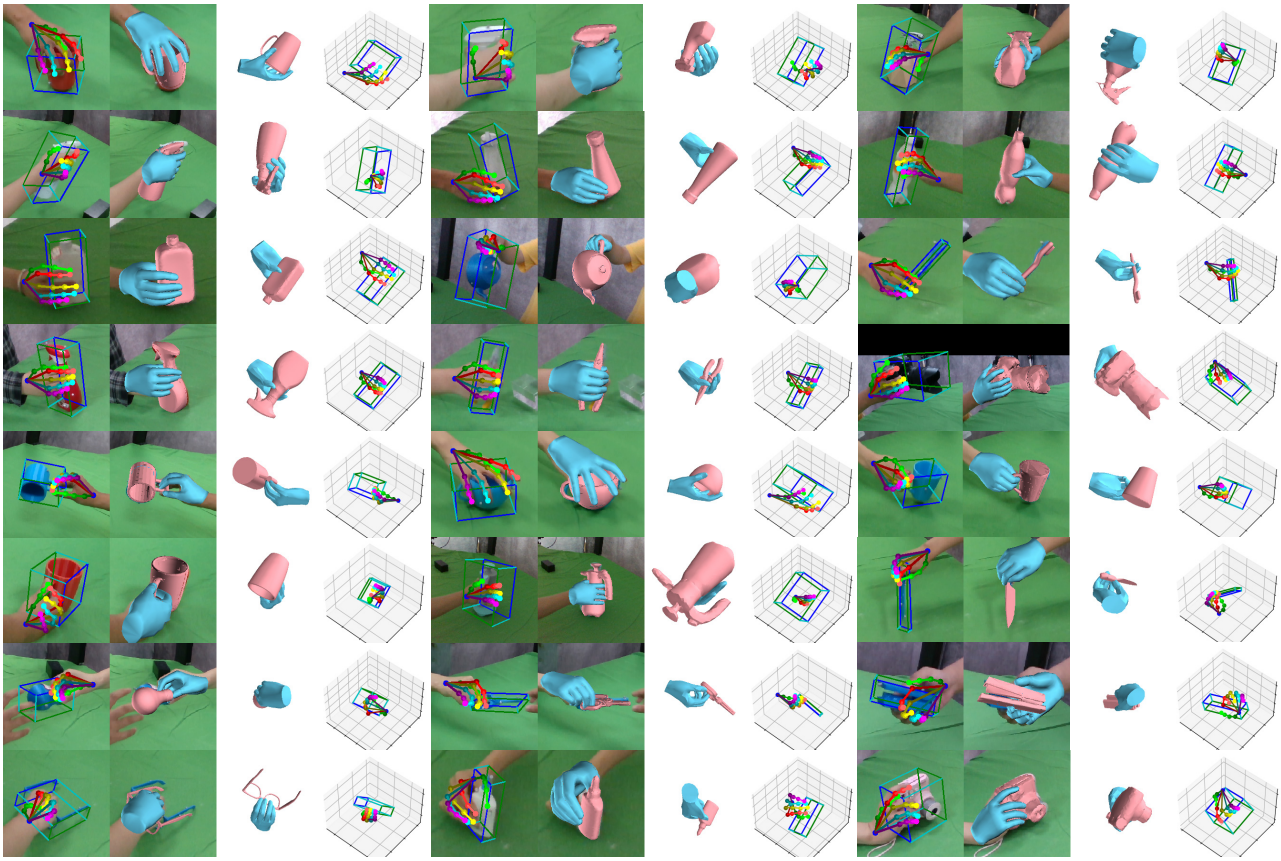

Figure 9. **More qualitative results** on HOPE task.

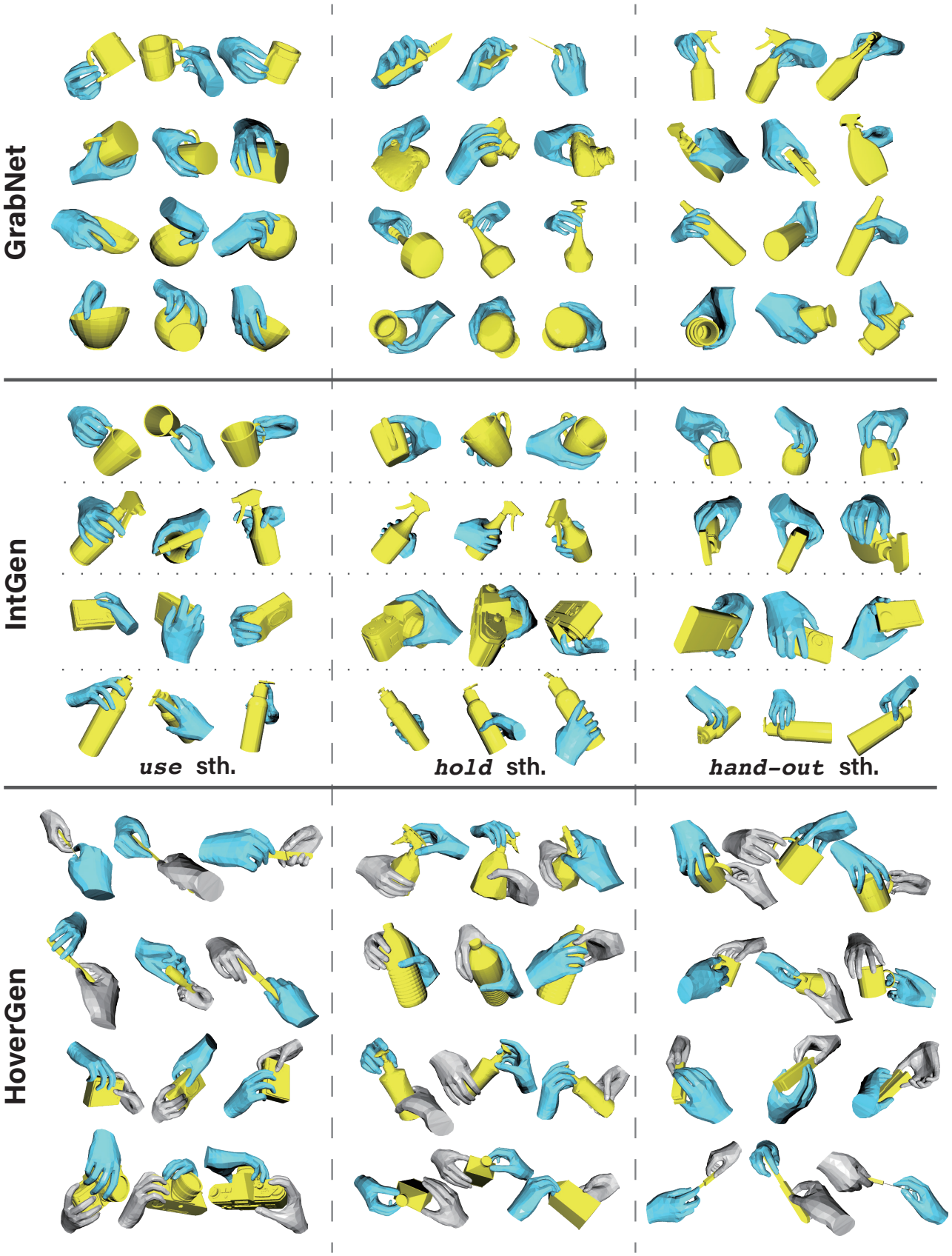

Figure 10. **More qualitative results** on GrabNet, IntGen and HoverGen predictions.