

LatentHOI: On the Generalizable Hand Object Motion Generation with Latent Hand Diffusion - Supplementary Material.

Muchen Li^{*1,2} Sammy Christen⁶ Chengde Wan⁵ Yujun Cai⁵
 Renjie Liao^{1,2,3} Leonid Sigal^{1,2,3,4} Shugao Ma⁵

Contents

| | |
|----------------------------------------------|----------|
| A More Experiment Details | 1 |
| A.1. Model implementation details | 1 |
| A.2. Dataset preparation details | 1 |
| A.3. Evaluation metric details | 2 |
| B More Experiment Results | 4 |
| B.1. Comparison to DiffH2O | 4 |
| B.2. More result on DexYCB dataset | 4 |
| B.3. More Qualitative Results | 4 |
| B.4. User Study statistics | 4 |

A. More Experiment Details

A.1. Model implementation details

GraspVAE For grasping VAE, we use MLP with two residual blocks for Encoder and Decoder respectively. For encoding the object pose condition, we use PointNet++ [5] with 3 blocks with two furthest point sampling and downsample layers and one final global pooling later.

LatentDiffusion For the Diffusion model, we follow Karunratanakul et al. [4] to use a 1D UNET as the backbone rather than the transformer backbone [7]. Concurrent work showed that [2] the 1D Unet performs on-par-with transformer for HOI. The hyperparameters used for the model can be found in Tab 1. We refer readers to [4] for detailed architecture design.

^{*} Corresponding author, work partially done during a internship at Meta Reality Lab, Redmond. Contact: muchenli@cs.ubc.ca

¹ University of British Columbia; ² Vector Institute for AI; ³ Canada CIFAR AI Chair; ⁴ NSERC CRC Chair; ⁵ Meta Reality Lab, Redmond; ⁶ ETH Zurich, Switzerland;

[†]Data collection, analysis, and experiments involving DexYCB were conducted at the University of British Columbia.

| Hyperparameters | Denoising Unet | Grasping VAE |
|---------------------------------|-----------------------------------|----------------------|
| Model Design | | |
| Type of layers | 2-layer Residual MLP | 1D Convolution layer |
| Number of Encoder Layers | 2 | 5 |
| Number of Decoder Layers | 2 | 5 |
| Hidden dimension | 512 | 1024 |
| Training Hyperparameters | | |
| Learning Rate | 2e-4 | 2e-4 |
| Optimizers | AdamW | |
| LR Scheduler | Cosine learning rate with warm up | |
| Weight Decay | 1e-4 | 1e-4 |
| Batch Size | 96 | 1024 |
| Number of Epochs Trained | 1250 | 20 |

Table 1. Training details for Denoising Unet and Grasping VAE

Mirroring the left-hand data For bi-manual hand motion, we mirror the left-hand data to the right-hand ones when training the grasping VAE, ensuring that we share VAE between right and left hand. To be more specific, in order to mirror the left-hand MANO parameters to the right-hand, one can select either the XZ or YZ plane and apply mirror translation to the rotation matrix and hand-root translation. During latent Diffusion training, the left-hand latent is calculated as the "pseudo" mirrored counterpart of the right-hand pose. When sampling, the pseudo right hand is decoded from the graspVAE and then mirrored back to obtain the left hand.

More detailed pipeline for sampling process In Fig 1 we show a detailed version of training and sampling pipeline of Figure 2 in the main paper.

A.2. Dataset preparation details

GRAB GRAB is a comprehensive full-body motion dataset that includes 3D shape and motion capture (mocap) pose sequences of 10 subjects interacting with 51 everyday objects of varying shapes and sizes. The dataset comprises a total of 1,334 sequences captured at 120 fps. The orig-

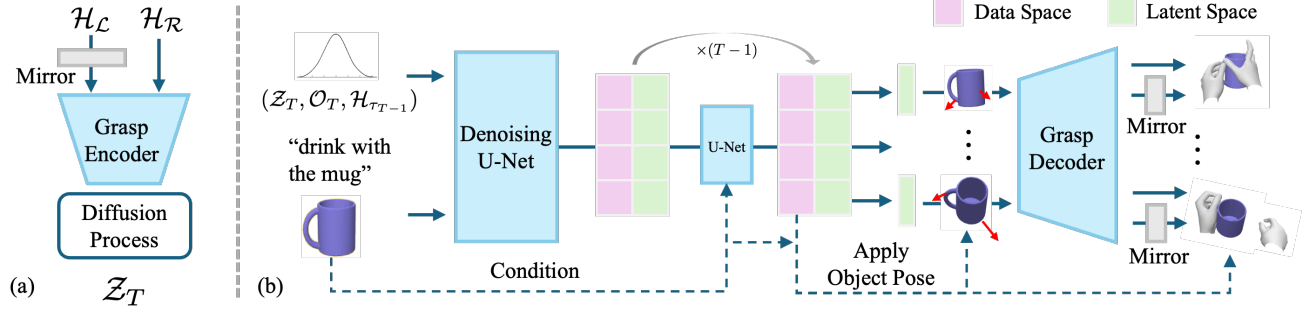


Figure 1. Detailed pipeline for training(a) and sampling (b)

inal GRAB dataset was recorded at 120fps with variable sequence lengths. Following Ghosh et al. [3], we first trim the sequences to include only the start and end of the interaction. We then downsample the sequences to 20fps and set the maximum duration to 8 seconds, resulting in a total of 160 frames. For motion sequences shorter than 8 seconds, we pad them to the target length using the last frame. In our experiments, the test split objects are an apple, a mug, a train, and an elephant. For all convenience, we ignore hand shape differences between subjects by using the mean shape provided by MANO [6]. This split allows 1.2k sequences for training over 47 objects and a total of 17 (text prompt, object) pairs over 4 novel objects for testing.

For VAE training, we split the motion sequences into frames and mirror the left hand to the right hand along with the interacting object. We observed that many off-contact poses for the left hand could lead to data imbalance during training, so we only retain left-hand data that are in contact with the object. This result to 1.2M frames for training VAE.

For encoding the intent into text descriptions, we use the simple action + object name setting (e.g., "pass the apple").

OakINK We use the virtual objects collected in Yang et al. [8]. Since there is no motion sequence provided by oakink, we use the same model that's trained on GRAB and test its generalize ability on oakink dataset. We first select 20 object categories that appear in the GRAB dataset (except for the driller, which appears only in DexYCB). For each class, we sample five diverse objects with different 3D shapes. To test on these novel objects, we match the corresponding intent for each object that appears in the GRAB dataset. This results in a total of 212 (text prompt, object) pairs. A visualization of all 3D shapes and their tested intents is shown in Figure 2 and Table 2. For training the,

DexYCB Compared to the GRAB dataset, DexYCB involves single-hand motion with fewer frames and a lower frame rate. We retain the reaching out and interaction phases and pad all grasping sequences to 96 frames. We also split

Table 2. Actions associated with various objects (updated)

| Object | Actions |
|----------------|--------------------------|
| binoculars | see, offhand |
| bowl | drink, pass |
| camera | take picture, browse |
| can | inspect, offhand |
| drill | pass, lift |
| eyeglasses | clean, wear |
| flashlight | on, offhand |
| fryingpan | cook, pass |
| gamecontroller | play, lift |
| hammers | use, pass |
| headphones | use, offhand |
| knife | chop, peel |
| lightbulb | pass, screw |
| mouse | use, offhand |
| mug | drink, toast |
| phone | call, pass |
| teapot | pour, lift |
| toothbrush | brush, lift |
| wineglass | drink, toast |
| waterbottle | open, shake, drink, pour |
| screwdriver | pass, lift |
| donut | eat, lift |

the train and test sets in the same manner as done for GRAB. Out of the 20 objects presented in DexYCB, we choose "banana," "foam brick," "master chef," and "mug" for the test set, with the remaining objects allocated to the training set. This results in 800 sequences for training and 200 sequences for testing.

A.3. Evaluation metric details

Interpenetration Volume per contact Unit (IVU) The Interpenetration Volume per Contact Unit is designed to quantify the extent of penetration relative to the degree of

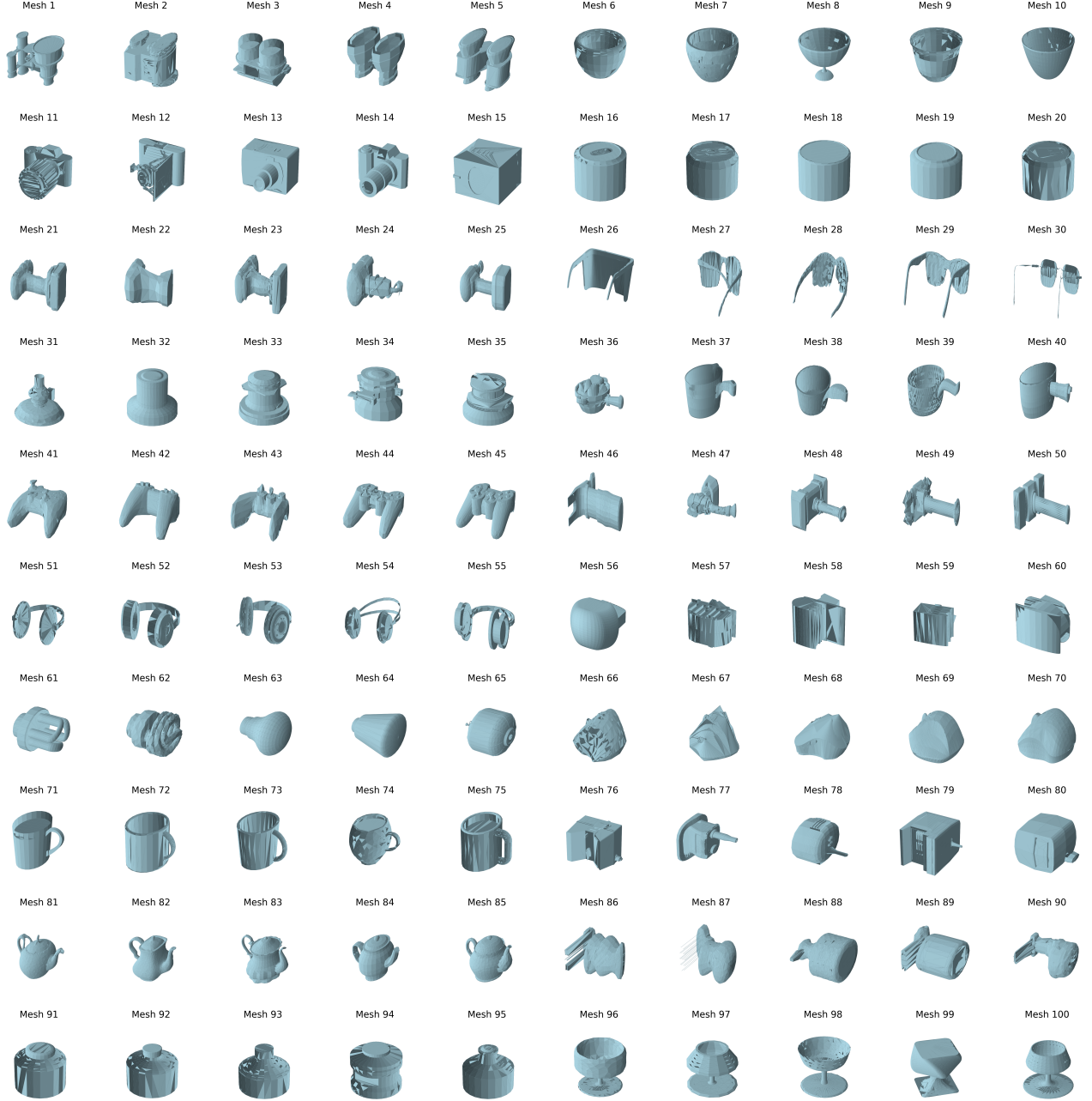


Figure 2. Our Novel Objects Split selected from [8] virtual objects.

contact.

$$IVU = \frac{\text{Interpenetration Volume (cm}^3\text{)}}{\text{Contact Region (cm}^2\text{)}}$$

Notably, we approximate the contact region by multiplying the Contact Ratio (CR)—defined as the proportion of hand vertices in contact with the object—by the mesh area.

Physical Plausibility Phy The Phy metric is designed to heuristically evaluate the plausibility of a grasp, based on the principle that contact forces must support the object when it is off the ground.

$$Phy = \frac{\sum_1^N \mathbf{1}(CR_i > 1\%)}{N}$$

A high Phy score represents a necessary but not sufficient condition for a physically plausible grasp. However, for eval-

Table 3. Comparison with DiffH2O on the Grab dataset, following their setup as described in [2]. Note that our implemented MDM* can be viewed as DiffH2O w/o canonical representation of hands and signed distance field for joints.

| Model | IV [cm ³] ↓ | ID [mm] ↓ | CR [%] ↑ | IVU ↓ |
|-------------|---------------------------|--------------|--------------|---------------|
| DiffH2o | 9.03 | 11.39 | 8.60 | 0.2576 |
| MDM* | 9.04 | 11.26 | 9.23 | 0.2403 |
| Ours | 8.36 | 11.14 | 14.76 | 0.1390 |

uation, we can justifiably conclude that a grasp is definitely not plausible if it has a low Phy score.

Sample Diversity (SD) and Overall Diversity (OD) Diversity is measured by per-pair L2 distance of motion sequences, for a set of motion sequence v_1, v_2, \dots, v_N

$$Diversity = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i}^N ||v_i - v_j||$$

Note that we followed previous work [2] to use this metric. However, this is a sub-optimal metric, primarily because it is not rotationally invariant. Even a small rotation of a trajectory can result in a large L2 distance from itself, despite representing the same motion sequence. It’s also shown in Table 1 of the main paper that both sample diversity and overall diversity correlate poorly to the quality of generated sequence.

B. More Experiment Results

B.1. Comparison to DiffH2O

Since diffh2o did not release code, we follow their interaction-only set up and unseen object split to compare our method. Under this setting, our mdm baseline exactly match the setup for diffh2o with two differences 1. hand object representation is not canonicalized to object 2. we don’t model the signed distance field for keypoints. We made sure that diffh2o, mdm and our method uses the same 1D Unet Backbone (please refer to table 6 in [2]). We find that our method exhibit less penetration while having higher contact rate.

B.2. More result on DexYCB dataset

Here we provide more result on DexYCB dataset. Since the training of our GraspVAE is decoupled from the training of latent diffusion, an additional advantage of our model is its ability to leverage widely available single-frame grasp data. To explore this, we conduct experiments using single-frame grasp data from the OakINK dataset [8]. Specifically, we incorporate all frames from the GraspVAE dataset along with single-frame data from the OakINK dataset to train our

Table 4. More quantative evaluation results on dexYCB dataset.

| Model | IV ↓ | ID ↓ | CR ↑ | IVU ↓ | Phy ↑ | SD ↑ | OD → |
|-----------------------|-------------|-------------|--------------|-------------|---------------|-------------|-------------|
| MLD[1] | 9.25 | 1.84 | 8.29 | 0.27 | 71.16 | 0.18 | 0.19 |
| MDM[7] | 7.78 | 2.10 | 8.87 | 0.18 | 86.22 | 0.13 | 0.13 |
| Ours | 7.70 | 2.01 | 11.98 | 0.13 | 88.52 | 0.13 | 0.12 |
| + OakInk single frame | 7.47 | 1.89 | 11.61 | 0.15 | 100.00 | 0.12 | 0.12 |

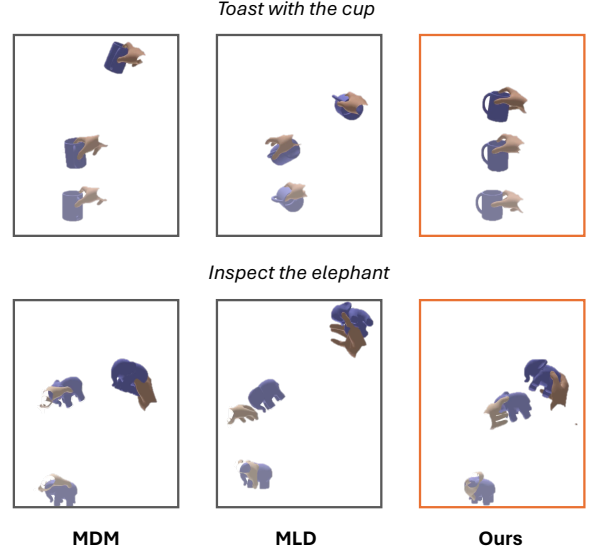


Figure 3. More result on GRAB dataset

GraspVAE. The results, presented in Table 4, demonstrate that training GraspVAE on single-frame data yields comparable penetration metrics while achieving higher physical plausibility.

B.3. More Qualitative Results

Here we show more qualitative results as sequences in Figure 3 5 6 7 8.

We also shown two examples of failure cases from OakINK dataset, as shown in Figure 9. The driller, absent from the training set, led to a failure in generating the appropriate “drill” motion. Although hammers of various shapes were present in the training set, the novel hammer in the test split was three times larger than the largest seen during training, resulting in a grasping position mismatch.

B.4. User Study statistics

We used Google Forms to conduct a user study, with an example of the interface shown in Figure 4. To ensure fairness, we randomly flipped each of the two-sided videos with a probability of 0.5. The user study included a total of 51 clips, with 17 clips sampled from GRAB test cases, 26 sampled from Oakink test cases, and 8 sampled from the DexYCB test cases.

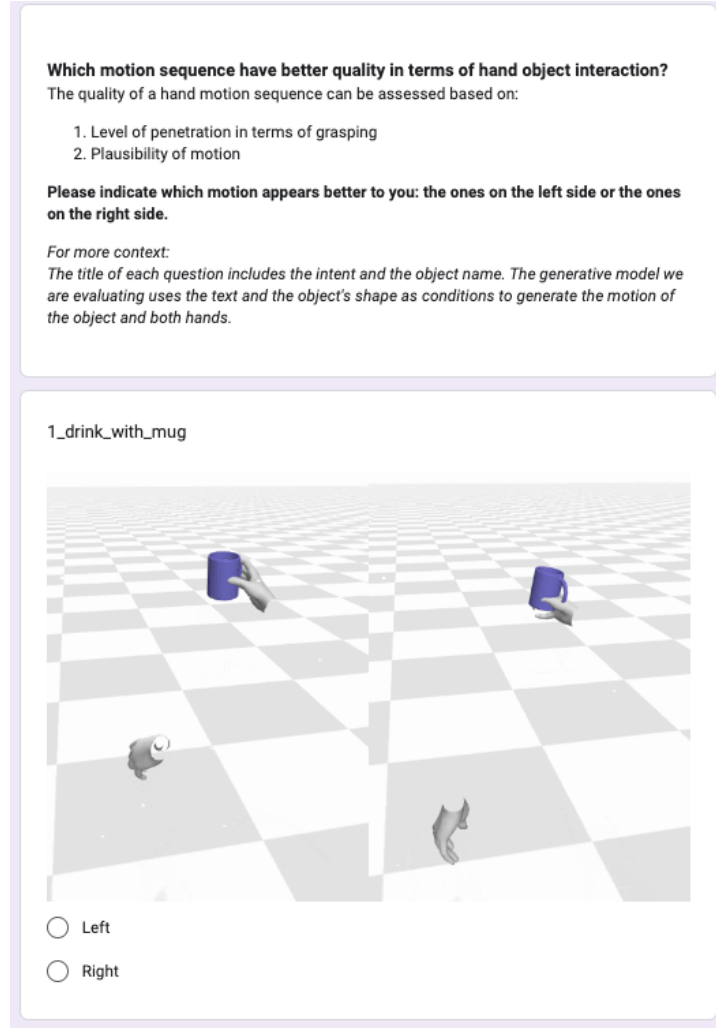


Figure 4. Screen shot of user study used in our experiments

References

- [1] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 4
- [2] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. *arXiv preprint arXiv:2403.17827*, 2024. 1, 4
- [3] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Eurographics*, 2023. 2
- [4] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 1
- [5] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1
- [6] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6), 2017. 2
- [7] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 4
- [8] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4

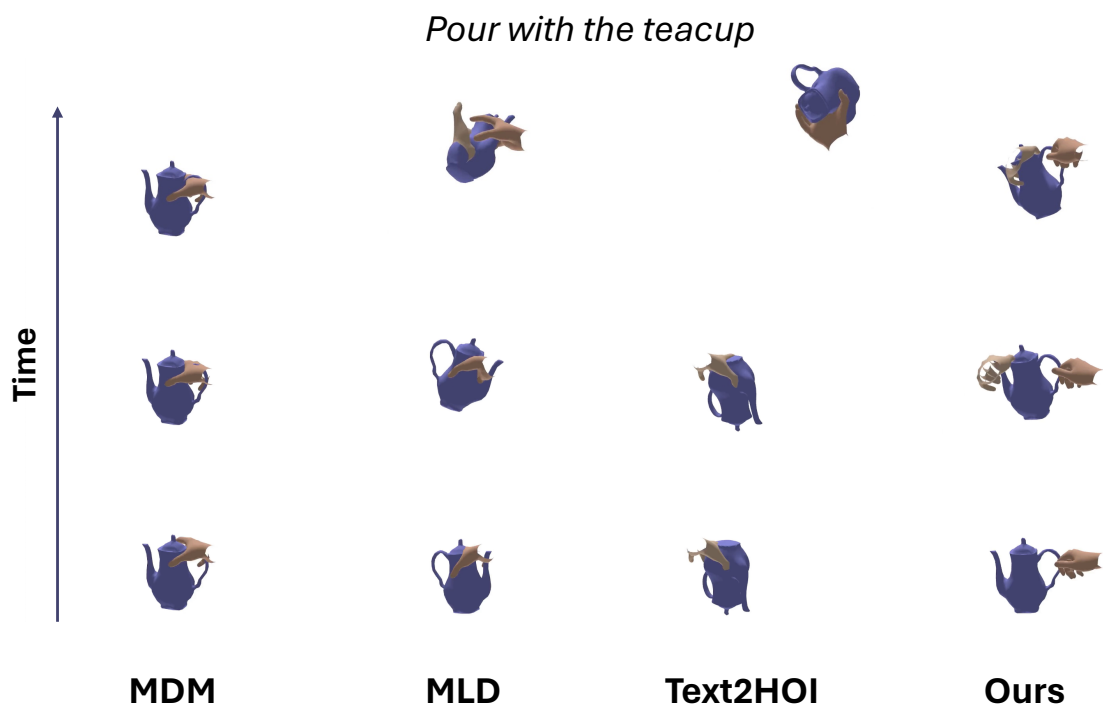


Figure 5. More result on Oakink dataset

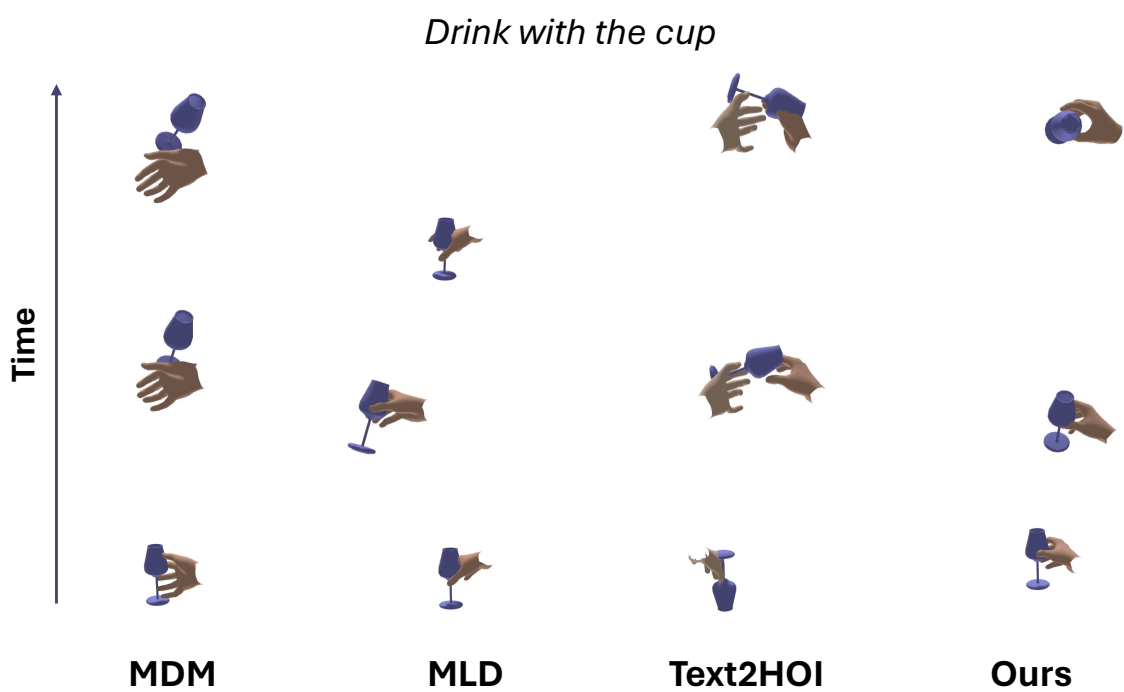


Figure 6. More result on Oakink dataset

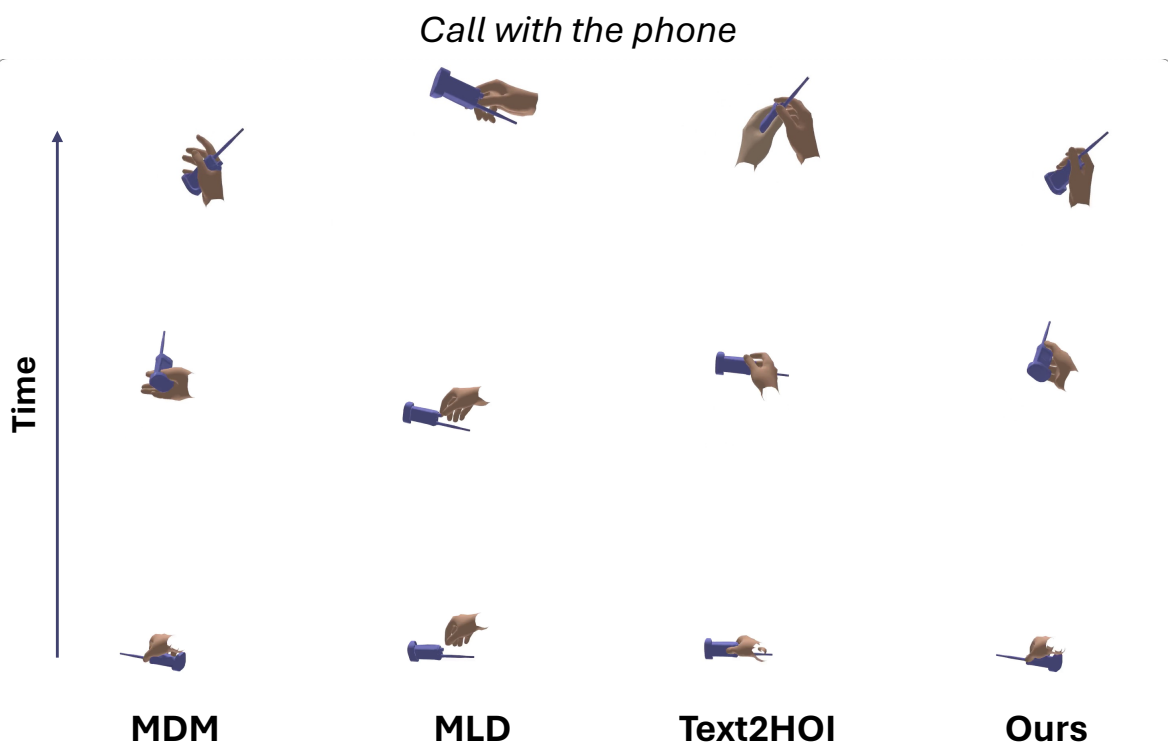


Figure 7. More result on Oakink dataset

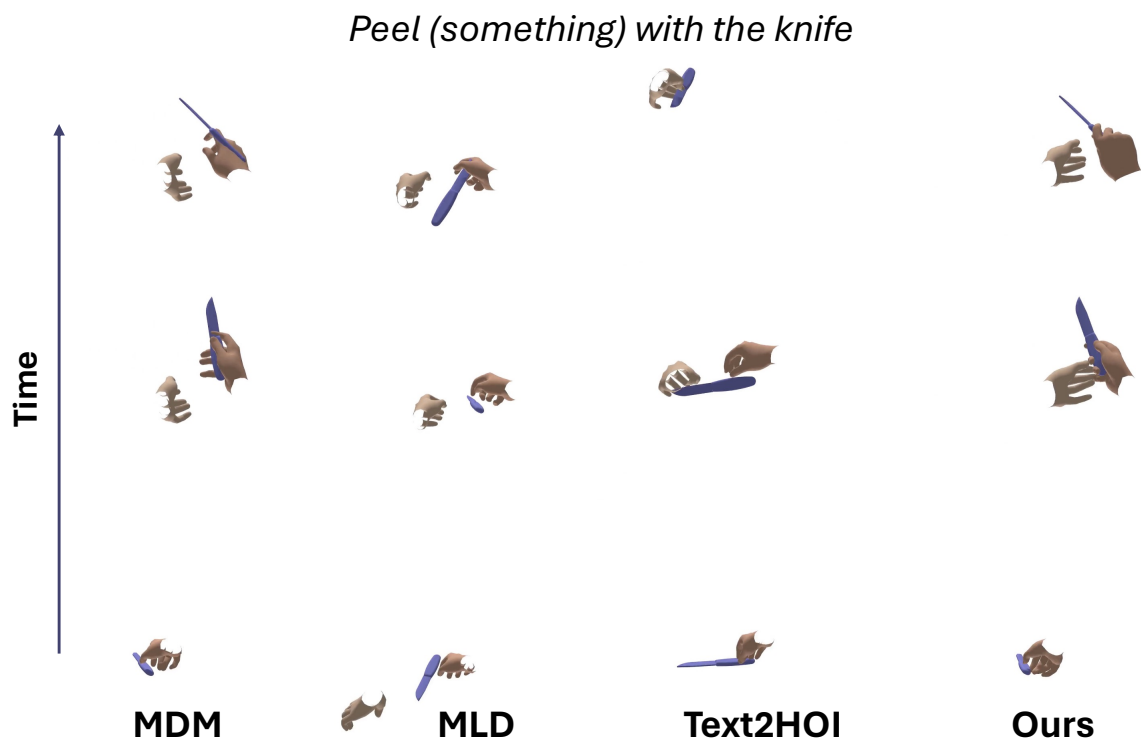


Figure 8. More result on Oakink dataset

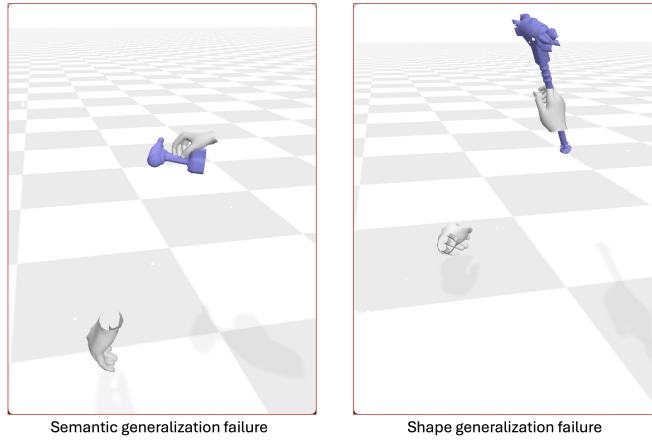


Figure 9. Failure cases in the OakInk test split. The prompts for the left and right are “*use the driller*” and “*use the hammer*”.