AffordPose: A Large-scale Dataset of Hand-Object Interactions with Affordance-driven Hand Pose

Juntao Jian Xiuping Liu Manyi Li Ruizhen Hu Jian Liu

ICCV 2023

Abstract

How humans interact with objects depends on the functional roles of the target objects, which introduces the problem of affordance-aware hand-object interaction. This requires a large number of human demonstrations for the learning and understanding of plausible and appropriate hand-object interactions. This work presents AffordPose, a large-scale dataset of hand-object interactions with affordance-driven hand poses. The dataset provides fine-grained, part-level affordance labels, segmentation for each object, and corresponding interaction hand poses such as twist, handle-grasp, and press.

Additionally, this paper designs a functional-driven hand pose generation framework and uses the AffordPose dataset for training, enabling the generation of natural interaction gestures driven by the functional properties of the object. Based on AffordPose and the associated network framework, this work extends the original affordance labels to text instructions and implements interaction hand pose generation driven by open-vocabulary instructions.

Keywords: Hand pose generation, Affordance, Grasping.

1 Introduction

One of the long-standing goals of robotics is to imitate all kinds of human-centered interactions, especially hand-object interactions, ranging from general grasping to functional interactions such as unscrewing a cap or even tool usage [22, 30]. Performing appropriate hand-object interaction is a complicated decision-making process. The agents need to understand the functional role of the object, select the contacting location, and perform the specific hand pose to complete the task [1, 10, 50].

There has been a trend to develop deep learning solutions to predict diverse hand-object interactions. The researchers build hand-object interaction datasets, such as HO-3D [21], DexYCB [8], Obman [23], and train different networks [37, 39, 56, 58] to predict the hand poses for the given objects. However, these works only consider the general grasping task and focus on the stability of the generated hand poses, but overlook the semantic meaning of the hand-object interactions.

This paper takes a step further to study the hand-object interactions driven by part-level affordances, which provide fine-grained localizations and are generalizable among object categories. We first collect the specific part-level affordances on the objects, i.e. the hand-centered labels such as twist, pull, handle-grasp, and the

corresponding parts, instead of the general labels such as use or handover, then manually adapt the hand poses to complete the interaction tasks corresponding to these affordances. As shown in AffordPose dataset, the part-level affordances correspond to some common characteristics of the hand-object interactions even with different object categories, yet allows an extent of hand pose diversity, thus improving the understanding and prediction of hand-object interactions. AffordPose collects 26.7K manually annotated interactions, each including the 3D object shape, the part-level affordance label, and the parameters of the detailed hand configuration.

Besides, this paper also designs AffordPoseNet for affordance-oriented hand-object interaction generation, which aims to predict a possible hand pose, including intrinsic (joint configurations) and extrinsic (hand rotation and position) parameters, based on the input object and a given affordance label. Specifically, AffordPoseNet is a two-stage framework. The first stage uses a conditional VAE network [51] to predict coarse hand parameters, while the second stage utilizes Refine-Net to fine-tune the hand parameters. The outcome is the ability to predict specific hand poses based on a given affordance label.

The paper designs AffordPoseNet for the task of affordance-driven hand-object interaction hand pose generation, showing promising application prospects. However, due to dataset limitations, the objects and affordances are confined to the distribution within the dataset, restricting the ability to handle diverse tasks in an open-world setting. In recent years, with the popularity of language models like ChatGPT, there has been a surge in prediction tasks based on open language. Therefore, I have chosen to improve AffordPoseNet to enable interaction hand pose generation driven by open-vocabulary instructions.

2 Related works

2.1 Hand-Object Interaction Datasets.

It's a crucial problem to produce accurate and plausible hand poses to perform hand-object interactions. The researchers have developed different hand pose acquisition, reconstruction, and simulation methods to build large-scale datasets for tasks ranging from hand pose estimation [15, 40], and grasp synthesis [37, 44], to hand-object interaction generation [6, 11].

Some works focus on hand pose estimation from hand-object interaction observations, e.g. RGB, RGB-D, or video sequential inputs. Therefore, the essential datasets should contain a large amount of accurate hand-object interactions. For example, Bighand2.2 [68] collects million-scale hand poses by building a tracking system. Obman [23] utilizes a grasp optimizer to synthesize the hand poses while ensuring the stability of the grasping. HO-3D [21] and DexYCB [8] build the motion capture systems to collect the sequential frames with one or more RGB-D cameras and solves for the 3D hand and object poses to build their datasets. On the other hand, some related works, e.g. DexteriousGrasping [70] and DexGraspNet [59], synthesize for the robotic dexterous hand poses, to complete the grasping task. These constructed datasets provide a wide range of large-scale hand-object interactions for the learning of hand pose estimation from the input observations.

However, as pointed out in ContactGrasp [3], the hand-object interactions are not only stable but also functional. In other words, the hand-object interaction datasets should involve more human annotations, rather than being built automatically, to reveal how human use different objects. Some related works, e.g. ContactGrasp [3], ContactPose [5] and Contact2Grasp [31], require the annotators to specify the contact maps of each object and then optimize the hand poses via simulators. The contact map constrains the functional goal of the

interactions, while the simulator optimization ensures the physical feasibility of the hand poses. Alternatively, YCB-Affordance [11] requires the annotators to manually specify the hand position, hand pose, and grasp type of each object, and then transfer the grasps to the YCB scenes [62]. These datasets collect natural and realistic hand poses for different objects, which are necessary for the learning of hand-object interaction generation.

Some recent works investigate hand-object interactions with different intents. GRAB [52] captures the whole-body grasps for different interactions, e.g. eating a banana, drinking from a bowl, etc, which are classified into 4 different intents, i.e. use, pass, lift, and off-hand pass. Similarly, OakInk [64] collects affordance-aware and intent-oriented hand-object interactions. That is, the captured hand-object interactions are performed based on the semantic meaning of objects and the specified intents, including use, hold, lift-up, hand-out and receive. H2O [67] provides a particular benchmark for the human-human object handover analysis.

This work builds the dataset named AffordPose, which contains large-scale hand-object interactions with affordance-driven hand poses. Our collected data, termed as affordance-driven hand-object interactions, are performed with the guidance of part-level affordance labels such as twist, pull, handle-grasp, etc. It is different from the grasp type labels in the YCB-Affordance dataset [11] which only indicates different joint arrangements of the hands, or the intents in OakInk dataset [64] which only indicates the general task purpose regardless of the object categories. Although people may consider the object affordances while performing the interactions, i.e. affordance-aware, the corresponding affordance for each interaction is ambiguous and not explicitly specified. By contrast, our dataset contains fine-grained hand-object interactions equipped with the corresponding part-level affordance labeling, which reveals the influence of hand-centered affordances on the detailed arrangement of the hand poses and allows the affordance-oriented hand-object interaction generation.

2.2 Hand Grasping

Due to the high flexibility of dexterous hands, grasping tasks with dexterous hands have become increasingly significant in achieving human-like robotic manipulation [49] [18] [63] [47]. Currently, most related methods [36] [59] [55] [28] [24] [60] primarily aim to achieve stable and plausible grasps in various ways, without considering subsequent tasks. Several optimization-based approaches [24] [42] [33] [46] [36] [59] [55] utilize metrics such as force closure [13] and collision detection to achieve stable dexterous grasp synthesis. For example, Obman [24] utilizes GraspIt! [45] to synthesize feasible and collision-free grasps, while DexGraspNet [59] leverages a deeply accelerated differentiable force closure estimator to generate numerous stable grasp results. Although these methods do not rely on data-driven training, they often lack realism in the grasp poses.

Data-driven approaches [16] [38] [43] [61] [28] [35] [60] [63], leverage annotated datasets [24] [20] [9] [65] [26] [59] to learn and generate plausible grasp poses. GanHand [12] utilizes a multi-task GAN [48] architecture to identify the optimal grasp type from 33 classes [17] and refines a hand model for realistic grasp synthesis. Other more methods [53] [60] [28] [34] employ Variational Autoencoders (VAEs) [57] [29] to encode information, learn the data distribution, and sample grasp poses from the learned distribution. Recently, UniDexGrasp [63] [18] has achieved remarkable generalization ability by employing a conditional probability model along with curriculum learning. Additionally, with the success of diffusion models [25], some works [43] [61] [7] have attempted to employ a diffusion-based denoising process to facilitate grasp synthesis. It is worth noting that some methods, such as [19] [4] [27] [32] [54] [66] [41], focus on using contact maps [2] to guide grasp generation, making the generated grasps more natural and physically feasible due to the

prior knowledge from human-annotated contact maps. Besides, FunctionalGrasp [69] [71] uses touch codes to synthesize functional grasps. However, these methods primarily consider the geometric features of objects and do not account for the differences and characteristics of interactive tasks. As a result, they cannot generate grasps with different task intentions for the same object.

3 Method

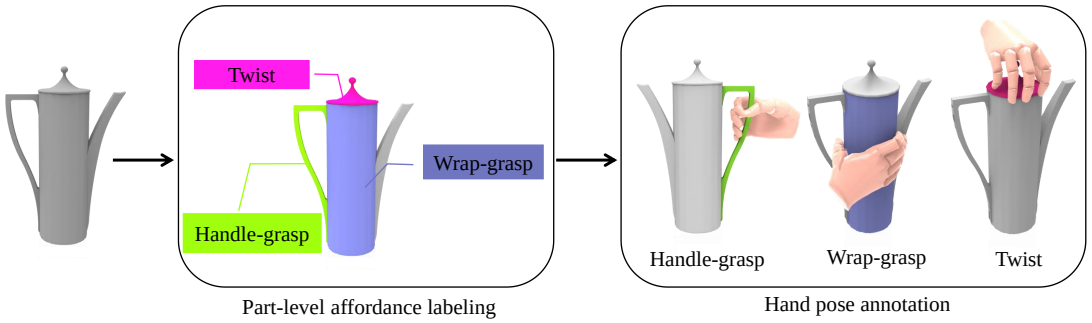

Figure 1. Dataset construction process for AffordPose. Firstly, annotating the part-level affordance labeling and then using it as guidance for the volunteers to adjust the hand pose annotations manually. [26]

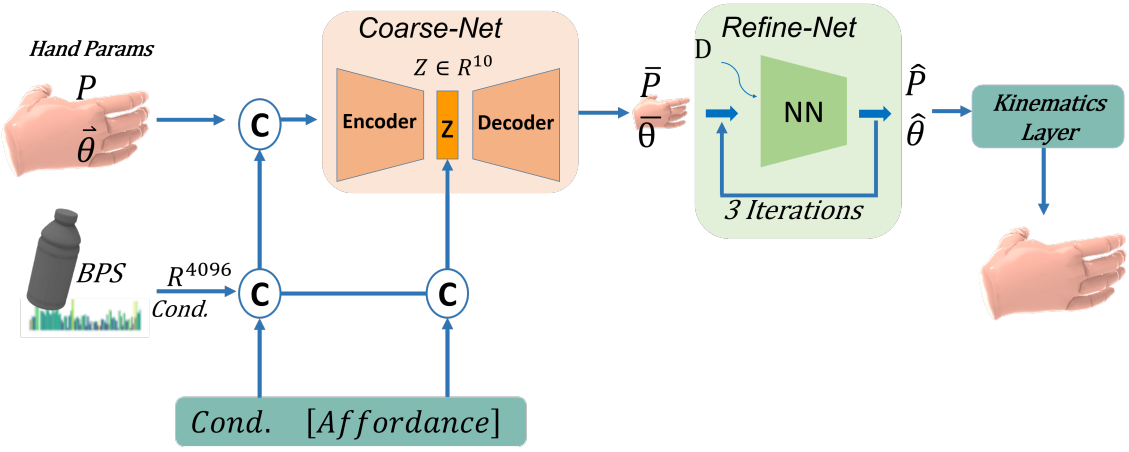

Figure 2. The network (AffordPoseNet) architecture of AffordPoseNet for affordance-oriented hand-object interaction generation. [26]

How humans interact with objects depends on the functional roles of the target objects, which introduces the problem of affordance-aware hand-object interaction. This requires a large number of human demonstrations for the learning and understanding of plausible and appropriate hand-object interactions. This work presents AffordPose, a large-scale dataset of hand-object interactions with affordance-driven hand poses. The dataset provides fine-grained, part-level affordance labels, segmentation for each object, and corresponding interaction gestures such as twist, handle-grasp, and press.

This paper also designs AffordPoseNet for affordance-oriented hand-object interaction generation, which aims to predict a possible hand pose, including intrinsic (joint configurations) and extrinsic (hand rotation and position) parameters, based on the input object and a given affordance label. Specifically, AffordPoseNet is a two-stage framework. The first stage uses a conditional VAE network [29] to predict coarse hand parameters, while the second stage utilizes Refine-Net to fine-tune the hand parameters. The outcome is the ability to predict specific hand poses based on a given affordance label.

3.1 AffordPoseNet

- **Data Pre-processing:** For training, we center each hand-object grasp sample at the centroid of the object and compute the $BPS \in \mathbb{R}^{4096}$ representation for the object, and the affordance labels are encoded as $A \in \mathbb{R}^{10}$, which are used for conditioning.
- **CoarseNet:** We pass the object shape BPS_o along with the initial MANO wrist rotation $\theta_{wrist} \in \mathbb{R}^{16}$ and translation $\gamma \in \mathbb{R}^6$ to the encoder $Q(Z|\theta_{wrist}, \gamma, BPS_o, A)$, which produces a latent grasp code $Z \in \mathbb{R}^{10}$. The decoder $P(\hat{\theta}, \hat{\gamma}|Z, BPS_o, A)$ then maps Z and BPS_o, A to the MANO parameters with full finger articulation, generating a 3D grasping hand.
- **RefineNet:** The grasps estimated by CoarseNet are plausible, but they can be refined for improved contact precision. For this, RefineNet takes as input the initial grasp $(\hat{\theta}, \hat{\phi})$ and the distances D from the MANO vertices to the object mesh. These distances are weighted according to the vertex contact likelihood learned from AffordPose. Then, RefineNet estimates the refined MANO parameters (θ', ϕ') in 3 iterative steps to produce the final grasp.
- **Kinematics Layer:** We need to construct a differentiable forward kinematics code based on the hand kinematics model, enabling a differentiable transformation between the hand parameters $(\hat{\theta}, \hat{\phi})$ and the hand mesh $M \in \mathbb{R}^{946}$.

4 Implementation details

4.1 Reproduction Description

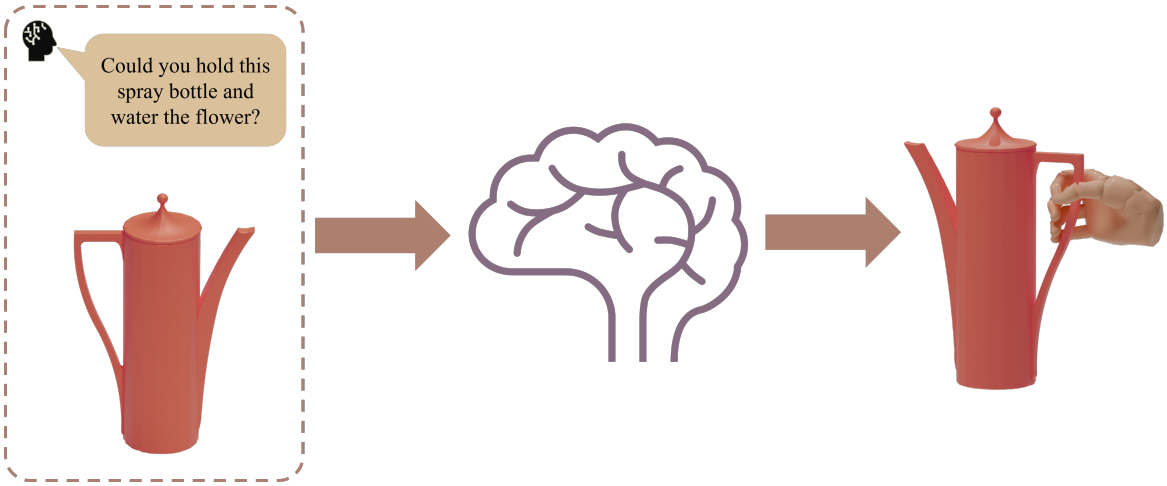

Figure 3. The pipeline of the grasping hand pose generation that is driven by the open-vocabulary instructions.

The paper designs AffordPoseNet for the task of affordance-driven hand-object interaction hand pose generation, showing promising application prospects. However, due to dataset limitations, the objects and affordances are confined to the distribution within the dataset, restricting the ability to handle diverse tasks in an open-world setting. In recent years, with the popularity of language models like ChatGPT, there has been a surge in prediction tasks based on open language. Therefore, I have chosen to improve AffordPoseNet to enable interaction hand pose generation driven by open-vocabulary instructions.

Open-vocabulary task

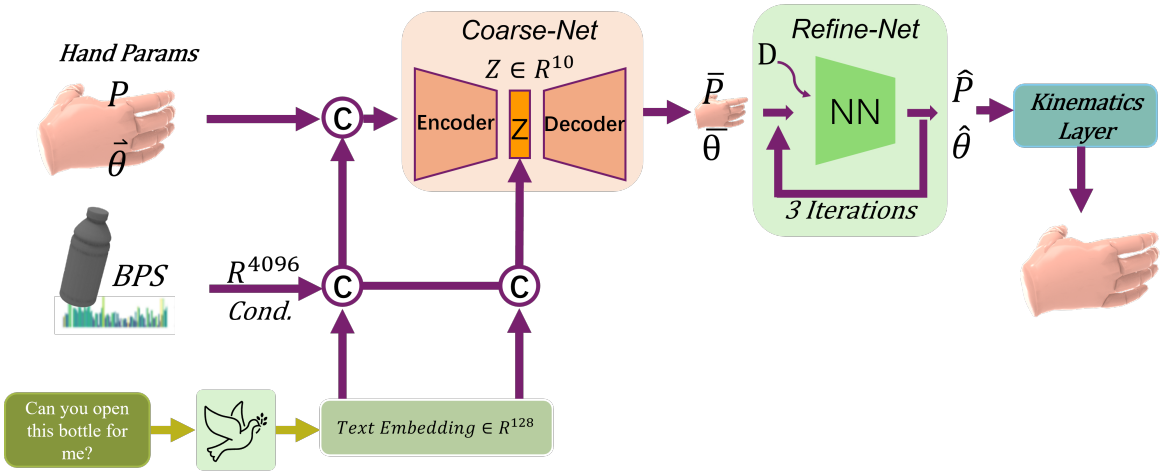

Figure 4. The AffordPoseNet-Bert pipeline of the grasping hand pose generation that is driven by the open-vocabulary instructions.

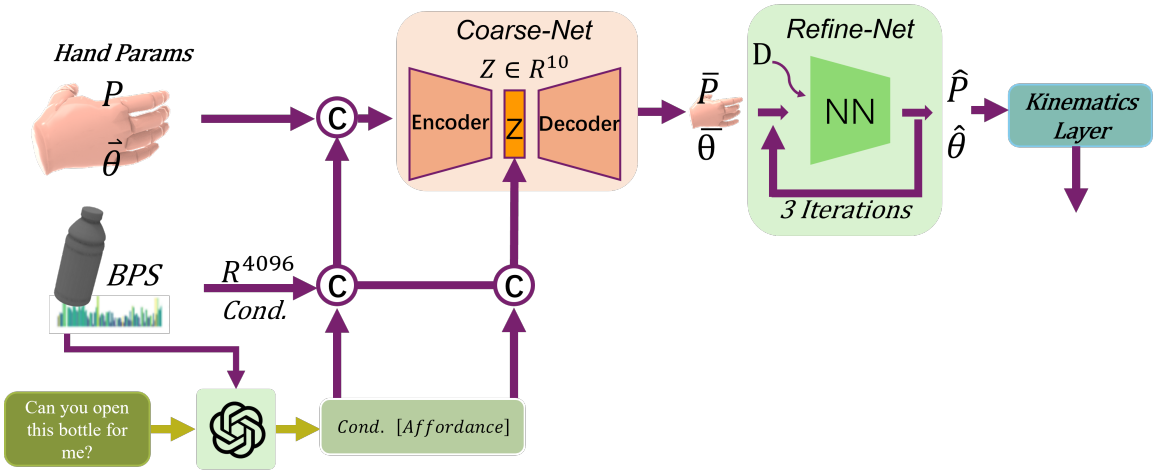

Figure 5. The AffordPoseNet-GPT pipeline of the grasping hand pose generation that is driven by the open-vocabulary instructions.

- **Text Instruction:** To achieve the task of generating interactive hand gestures guided by open-vocabulary instructions, we generate textual data for the grasping data in the AffordPose dataset, adding corresponding task instructions to each grasp. Specifically, we first highlight the interactive regions of each object in AffordPose, and then use the multimodal large language model ChatGPT-4 to generate task instructions that match the highlighted interactive region.

- **AffordPoseNet-Bert:** We encode the textual instructions Text using BERT [14] to obtain a $T \in \mathbb{R}^{128}$, which serves as the conditioning input for the original condition A.

- **AffordPoseNet-GPT:** In addition to using Bert to handle open-vocabulary instructions, one feasible approach is to leverage the powerful commonsense reasoning ability of multimodal large language models(e.g. GPT-4o) to parse the instructions, mapping the open instructions to the affordance labels defined in the dataset. This allows for direct hand pose prediction using the original AffordPoseNet.

4.2 Comparing with the released source codes

- AffordPose is my own work, and no related source code has been made available previously.
- AffordPoseNet-Bert and AffordPoseNet-GPT both are implemented by myself, including the dataset of the text instructions and the prediction of hand pose driven by open-vocabulary instructions based on Bert and GPT-4o API.

4.3 Experimental environment setup

The experiments were conducted on a machine with the following hardware configuration: an NVIDIA GeForce RTX 3060 GPU. The operating system used was Ubuntu 20.04. For the software environment, Python 3.8 was employed as the primary programming language, with key libraries including Pytorch3d, Numpy, PyTorch 1.9, Chamfer-distance, MANO, BPS-torch, Mesh.

5 Results and analysis

5.1 AffordPoseNet

Figure 6 shows the qualitative evaluations of the two experiments. As expected, in the first row, although GrabNet [51] is trained with various hand poses for different affordances, it can only produce roughly reasonable but similar hands for each object. Taking the bottle object (the 2nd and 3rd results in the top row) as an example, the predicted hands are similar even if we sample different random vectors from the Gaussian distribution to generate the results. Actually, when one object has several affordances, the predicted hand from GrabNet [51] often contacts the object in the middle of the related parts, with the hand pose corresponding to the most frequent affordance, i.e. wrap grasp for the bottle cases in Figure 6. By contrast, the AffordPoseNet is able to predict the distinctive hand poses for the specified affordance, justifying the effectiveness of the affordance labeling in guiding hand-object interactions. Most of the results of AffordPoseNet match with the input affordance condition, while the GrabNet baseline often generates similar hand poses when the input object has multiple affordances.

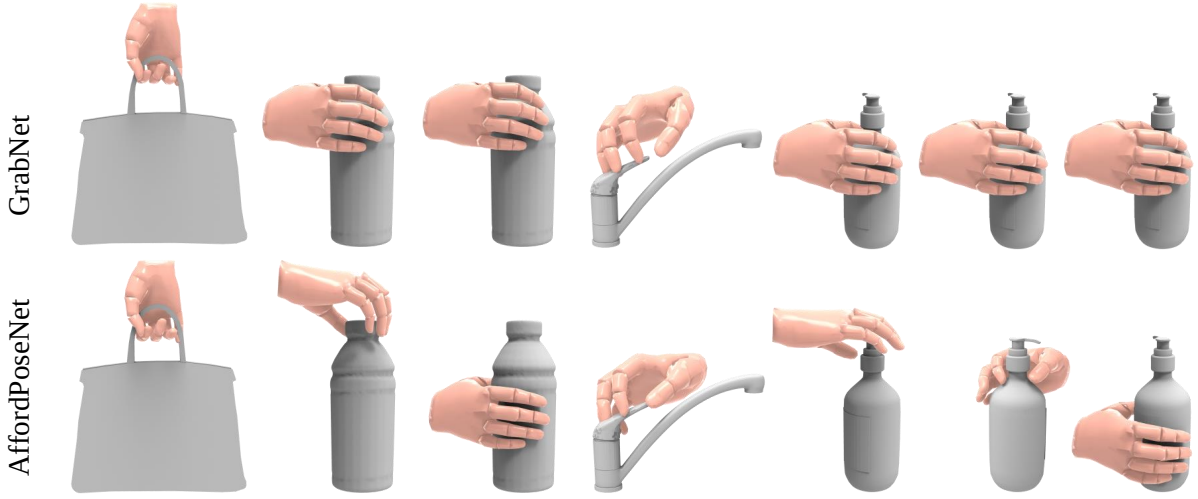

Figure 6. Experimental results of AffordPoseNet

	Solid.Intsec.Vol(cm^3)↓	Penet.Depth(cm)↓	Sim.Dis(cm)↓	Part Acc.(%)
AffordPoseNet [26]	6.20	1.42	10.57	84.8
AffordPoseNet-Bert	8.76	1.44	10.58	75.6
AffordPoseNet-GPT	6.77	1.64	10.66	81.0

Table 1. Quantitative comparison. AffordPoseNet-Bert and AffordPoseNet-GPT use BERT and GPT to handle task instructions as conditions respectively.

5.2 Open-vocabulary Task

As shown in Figure 7, the experimental results of AffordPoseNet-BERT are visualized. From the results, we can observe that by encoding the task text using BERT as a conditioning input, the open-vocabulary instruction-based hand pose generation task can be handled to some extent. However, as seen on the right side of Figure 7, there are some poor examples, which may be due to significant differences in object categories and shapes.

The quantitative results in Table 1 show that although simple encoding can generate suitable hand poses for most objects and tasks, compared to the original AffordPoseNet, this improvement leads to a decline in quality across all metrics.

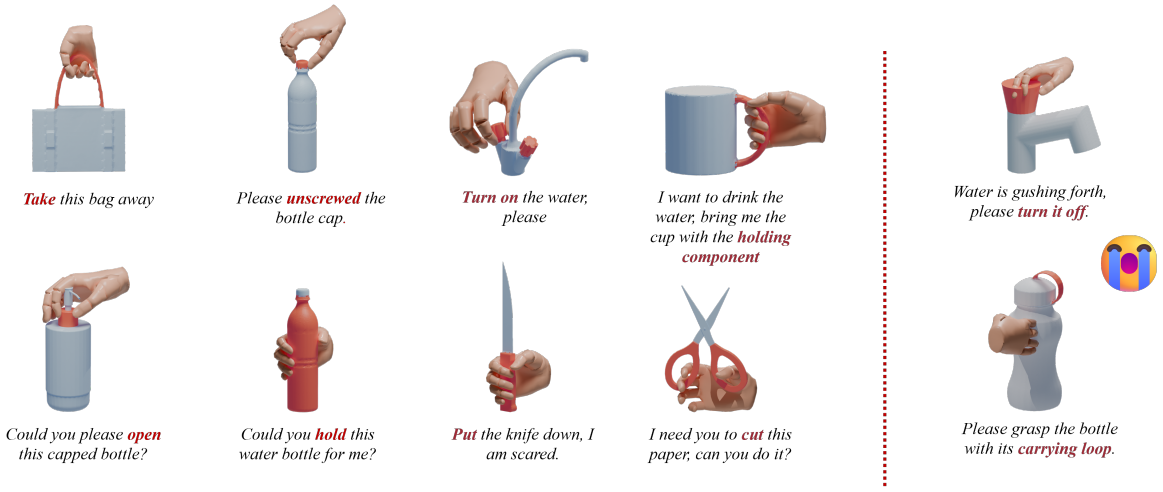

Figure 7. Experimental results of Open-vocabulary task based on AffordPoseNet-Bert

6 Conclusion and future work

We present AffordPose, a large-scale dataset of hand object interactions with affordance-driven hand poses. Our data analysis reveals how affordance affects the detailed configurations of the hand poses to complete the interactions. The additional affordance labeling helps to form the fine-grained hand-object interactions: the hand poses corresponding to the same affordance exhibit some distinctive characteristics as well as a certain degree of diversity. The effectiveness of our dataset is observed in the affordance-oriented interaction hand pose generation.

Additionally, we augment the dataset with textual instructions. By leveraging BERT and multimodal large language models on top of AffordPoseNet, we enable open-vocabulary, task-driven hand pose generation.

Since simple encoding of textual data cannot handle the complexity and variety of objects, we plan to focus on open-world settings in the future, aiming for generalization across diverse tasks and objects, as shown in Figure 8.

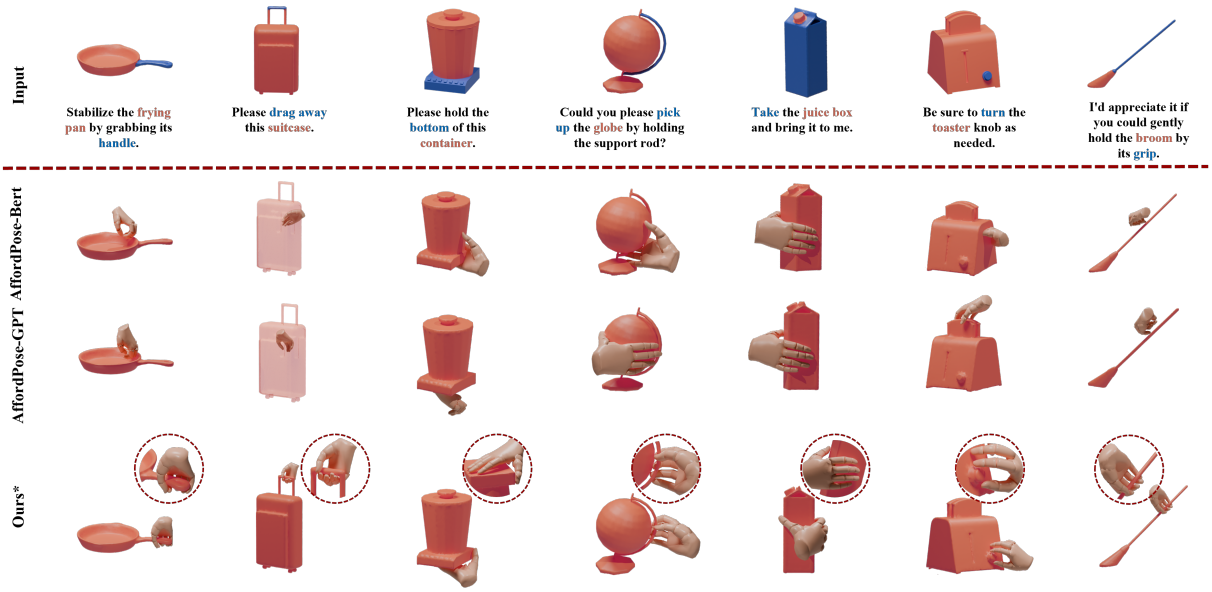

Figure 8. Experimental results of Open-worlds task.

References

[1] Paola Ard'on, Éric Pairet, Katrin Solveig Lohan, Subramanian Ramamoorthy, and Ronald P. A. Petrick. Affordances in robotic tasks - a survey. *ArXiv preprint arXiv:2004.07400*, 2020.

[2] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8701–8711, 2019.

[3] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393, 2019.

[4] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393, 2019.

[5] 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 *European Conference on Computer Vision*, 2020.

[6] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021.

[7] Xiaoyun Chang and Yi Sun. Text2grasp: Grasp synthesis by text prompts of object grasping parts. *ArXiv*, abs/2404.15189, 2024.

[8] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9049, 2021.

[9] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9049, 2021.

[10] Fu-Jen Chu, Ruinian Xu, and Patricio A. Vela. Learning affordance segmentation for real-world robotic manipulation via synthetic images. *IEEE Robotics and Automation Letters*, 4:1140–1147, 2019.

[11] Enric Corona, Albert Pumarola, G. Alenyà, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5040, 2020.

[12] Enric Corona, Albert Pumarola, G. Alenyà, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5040, 2020.

[13] Hongkai Dai, Anirudha Majumdar, and Russ Tedrake. *Synthesis and Optimization of Force Closure Grasps via Sequential Semidefinite Programming*, page 285–305. Jan 2018.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[15] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J. Crandall. Hope-net: A graph-based model for hand-object pose estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6607–6616, 2020.

[16] Haonan Duan, Peng Wang, Yayu Huang, Guangyun Xu, Wei Wei, and Xiao Shen. Robotics dexterous grasping: The methods based on point cloud and deep learning. *Frontiers in Neurorobotics*, 15, 2021.

[17] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on Human-Machine Systems*, 46:66–77, 2016.

[18] Haoran Geng and Yun Liu. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3868–3879, 2023.

[19] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. Contactopt: Optimizing contact to improve grasps. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1471–1481, 2021.

[20] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2019.

[21] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2020.

[22] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Comput. Surv.*, 54(3), 2021.

[23] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11799–11808, 2019.

[24] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11799–11808, 2019.

[25] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

[26] Ju Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14667–14678, 2023.

[27] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11087–11096, 2021.

[28] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. *2020 International Conference on 3D Vision (3DV)*, pages 333–344, 2020.

[29] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[30] Mike Land, Neil Mennie, and Jennifer M. Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28:1311 – 1328, 1999.

[31] Haoming Li, Xinzhuo Lin, Yang Zhou, Xiang Li, Yuchi Huo, Jiming Chen, and Qi Ye. Contact2grasp: 3d grasp synthesis via hand-object contact constraint. *ArXiv preprint arXiv:2210.09245*, 2022.

[32] Haoming Li, Xinzhuo Lin, Yang Zhou, Xiang Li, Yuchi Huo, Jiming Chen, and Qi Ye. Contact2grasp: 3d grasp synthesis via hand-object contact constraint. In *International Joint Conference on Artificial Intelligence*, 2022.

[33] Jiawei Li, Hong Liu, and Hegao Cai. On computing three-finger force-closure grasps of 2-d and 3-d objects. *IEEE Trans. Robotics Autom.*, 19:155–161, 2003.

[34] Kailin Li, Jingbo Wang, Lixin Yang, Cewu Lu, and Bo Dai. Semgrasp: Semantic grasp generation via language aligned discretization. *ArXiv*, abs/2404.03590, 2024.

[35] Kelin Li, Nicholas Baron, Xianmin Zhang, and Nicolás Rojas. Efficientgrasp: A unified data-efficient learning to grasp method for multi-fingered robot hands. *IEEE Robotics and Automation Letters*, 7:8619–8626, 2022.

[36] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8068–8074, 2022.

[37] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Generating grasp poses for a high-dof gripper using neural networks. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1518–1525, 2019.

[38] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Generating grasp poses for a high-dof gripper using neural networks. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1518–1525, 2019.

[39] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Deep differentiable grasp planner for high-dof grippers. *ArXiv preprint arXiv:2002.01530*, 2020.

[40] Shao-Wei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14682–14692, 2021.

[41] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20552–20563, 2023.

[42] Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *IEEE Robotics and Automation Letters*, 7:470–477, 2021.

[43] Jiaxin Lu, Hao Kang, Haoxiang Li, Bo Liu, Yiding Yang, Qixing Huang, and Gang Hua. Ugg: Unified generative grasping. *ArXiv*, abs/2311.16917, 2023.

[44] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6169–6176, 2021.

[45] Andrew T. Miller and Peter K. Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11:110–122, 2004.

[46] Jean Ponce, Steve Sullivan, Attawith Sudsang, Jean-Daniel Boissonnat, and Jean-Pierre Merlet. On computing four-finger equilibrium and force-closure grasps of polyhedral objects. *The International Journal of Robotics Research*, 16:11 – 35, 1997.

[47] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *ArXiv*, abs/1709.10087, 2017.

[48] Mathew Salvaris, Danielle Dean, and Wee Hyong Tok. *Generative Adversarial Networks*, page 187–208. Jan 2018.

[49] Qijin She, Ruizhen Hu, Juzhan Xu, Min Liu, Kai Xu, and Hui Huang. Learning high-dof reaching-and-grasping via dynamic representation of gripper-object interaction. *ACM Transactions on Graphics (TOG)*, 41:1 – 14, 2022.

[50] Hyun Oh Song, Mario Fritz, Daniel Goehring, and Trevor Darrell. Learning to detect visual grasp affordance. *IEEE Transactions on Automation Science and Engineering*, 13:798–809, 2016.

[51] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020.

[52] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020.

[53] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision*, 2020.

[54] Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ale Leonardis, Feng Zheng, and Hyung Jin Chang. S2contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. *ArXiv*, abs/2208.00874, 2022.

[55] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis fornbsp;multi-fingered hands. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, page 201–221, Berlin, Heidelberg, 2022. Springer-Verlag.

[56] Dylan Turpin, Liquang Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven J. Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. *ArXiv preprint arXiv:2208.12250*, 2022.

[57] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *ArXiv*, abs/1711.00937, 2017.

[58] Jacob Varley, Jonathan Weisz, Jared Weiss, and Peter K. Allen. Generating multi-fingered robotic grasps via deep learning. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4415–4420, 2015.

[59] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *ArXiv preprint arXiv:2210.02697*, 2022.

[60] Wei Wei, Daheng Li, Peng Wang, Yiming Li, Wanyi Li, Yongkang Luo, and Jun Zhong. Dvgg: Deep variational grasp generation for dextrous manipulation. *IEEE Robotics and Automation Letters*, 7:1659–1666, 2022.

[61] Zehang Weng, Haofei Lu, Danica Kragic, and Jens Lundell. Dexdiffuser: Generating dexterous grasps with diffusion models. *ArXiv*, abs/2402.02989, 2024.

[62] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. 2018.

[63] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, Tengyu Liu, Li Yi, and He Wang. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4737–4746, 2023.

[64] 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. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20921–20930, 2022.

[65] 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. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20921–20930, 2022.

[66] 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. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11077–11086, 2020.

[67] Ruolin Ye, Wenqiang Xu, Zhendong Xue, Tutian Tang, Yanfeng Wang, and Cewu Lu. H2o: A benchmark for visual human-human object handover analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15762–15771, 2021.

[68] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4866–4874, 2017.

[69] Yibiao Zhang, Jinglue Hang, Tianqiang Zhu, Xiangbo Lin, Rina Wu, Wanli Peng, Dongying Tian, and Yi Sun. Functionalgrasp: Learning functional grasp for robots via semantic hand-object representation. *IEEE Robotics and Automation Letters*, 8:3094–3101, 2023.

[70] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15721–15731, 2021.

[71] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15721–15731, 2021.