Robo-ABC : Affordance Generalization Beyond Categories via Semantic Correspondence for Robot Manipulation

Yuanchen Ju[1,2*] Kaizhe Hu[1,2,5*] Guowei Zhang[3,2] Gu Zhang[1,4,2]
Mingrun Jiang[2] Huazhe Xu[2,1,5†]

[1] Shanghai Qi Zhi Institute
[2] IIIS, Tsinghua University
[3] School of Software, Tsinghua University
[4] Shanghai Jiao Tong University
[5] Shanghai AI Lab
https://tea-lab.github.io/Robo-ABC/
juuycc0213@gmail.com, hkz22@mails.tsinghua.edu.cn

Abstract. Enabling robotic manipulation that generalizes to out-of-distribution scenes is a crucial step toward the open-world embodied intelligence. For human beings, this ability is rooted in the understanding of semantic correspondence among different objects, which helps to naturally transfer the interaction experience of familiar objects to novel ones. Although robots lack such a reservoir of interaction experience, the vast availability of human videos on the Internet may serve as a resource, from which we extract an affordance memory of contact points. Inspired by the natural way humans think, we propose Robo-ABC: when confronted with unfamiliar objects that require generalization, the robot can acquire affordance by retrieving objects that share visual and semantic similarities from the memory, then mapping the contact points of the retrieved objects to the new object. While such correspondence may present formidable challenges at first glance, recent research finds it naturally arises from pre-trained diffusion models, enabling **affordance mapping even across disparate categories**. Through the Robo-ABC framework, robots can generalize to manipulate out-of-category objects in a zero-shot manner without any manual annotation, additional training, part segmentation, pre-coded knowledge, or viewpoint restrictions. Quantitatively, Robo-ABC significantly enhances the accuracy of visual affordance inference by a large margin of **28.7%** compared to state-of-the-art (SOTA) end-to-end affordance models. We also conduct real-world experiments of cross-category object-grasping and achieve a success rate of **85.7%**, proving Robo-ABC's capacity for real-world tasks.

Keywords: Robot Perception · Affordance Inference · Cross-Category Generalization · Semantic Correspondence · Robot Manipulation

* Egual contribution.
† Corresponding author: Huazhe Xu (huazhe_xu@mail.tsinghua.edu.cn).

1 Introduction

Imagine a future where robots assist humans to proficiently accomplish a broad spectrum of daily tasks. The fundamental challenge lies in empowering robots to find interaction strategies with both familiar and novel objects. We humans instinctively have such abilities by generalizing the affordance [19] to unseen objects through semantic mapping [8]. For example, we may figure out how to grab a badminton racket by recalling the experience of wielding a tennis racket, or how to open a microwave oven that partially resembles a cabinet.

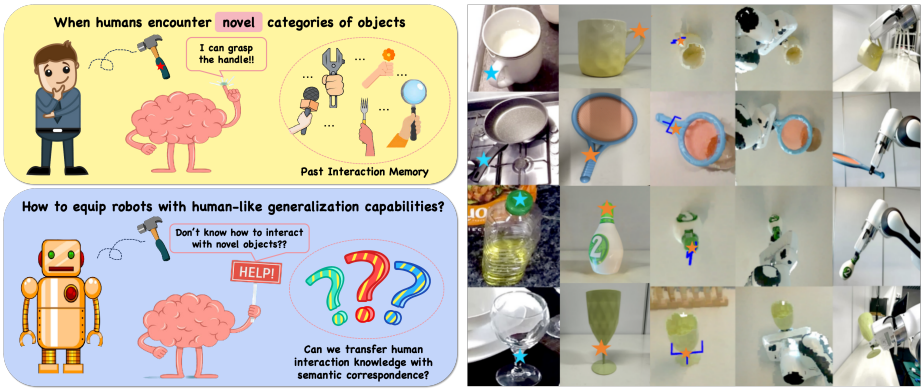


Fig. 1: Overview. The **left** illustrates the key insight of Robo-ABC (★ represents the contact point). Our goal is to endow robots with the same semantic correspondence ability as humans, which can generalize the object affordance across categories in manipulation tasks. The columns on the **right** in order, are source images (★ represents contact points which are extracted from human videos), corresponding attention maps on the target images (★ represents inferred contact points on unseen objects), grasp poses (Grasp poses are represented by ⊔, which are generated according to the contact points ★), point cloud during grasping, and the final successful grasping results.

However, this ability is not innate to robots. A key challenge is to obtain such interaction experience and extract generalizable information for robot manipulation. Fortunately, there is a wealth of egocentric human-object interaction videos [9,10,21] available on the Internet. These videos provides valuable insights into complex interactions, as well as the temporal and motion contexts of objects. Previous works have explored a variety of methods for learning object affordances from videos, such as extracting feature embeddings from videos [14], or automatically collecting pseudo-ground-truth labels for end-to-end training [3,4,35]. While existing methods can predict affordance for familiar objects, they struggle to generalize to unseen objects.

We aim to effectively and efficiently generalize affordance beyond object categories. To this end, we propose a general framework, Robo-ABC, that can extract object interaction experience from human videos and transfer it to novel objects. First, we extract the interaction experiences of objects from human videos and store them in an affordance memory. Second, in the face of a novel object (*e.g.*, a

screwdriver), we retrieve objects (*e.g.*, a knife) similar to the target object from the memory based on visual and semantic similarity. The most intriguing aspect of our method is the third step, where we employ the emergent semantic correspondence ability from the diffusion models [24, 36, 53, 66] to map the retrieved contact points to the novel object. We find this procedure powerful enough to transfer affordance beyond multiple object categories. Finally, we use the obtained affordance points to select from the grasping position prior provided by AnyGrasp [13] and deploy on a real robot to complete the manipulation tasks.

We conduct a comprehensive evaluation of Robo-ABC's generalization ability under different settings. We evaluate the zero-shot generalization ability with other end-to-end methods, and the affordance prediction success rate of Robo-ABC significantly increases by 28.7% compared to state-of-the-art (SOTA) end-to-end methods, demonstrating the strong generalization capability to novel objects and categories. We also demonstrate Robo-ABC's potential to generalize the affordance of one source object to objects that span a large category gap. Lastly, after deploying the Robo-ABC on a real robot using AnyGrasp [13], we show that our method can provide accurate affordance guidance for grasping in open-world, cross-view, clustered, and cross-category settings. Robo-ABC reaches a prominent success rate of 85.7% over 7 object categories.

Our contributions can be summarized as follows:

1) We propose a novel framework, Robo-ABC, to extract object interaction experience from human videos and transfer it to novel objects. Robo-ABC is the first robotic system to utilize foundation vision models for generalizing affordances across object categories with no need for annotation, additional training, or pre-coded knowledge of any kind.

2) We conduct exhaustive experiments on the performance of different semantic correspondence techniques on our pipeline and propose a simple yet effective method to overcome the sensitivity of image orientation for those methods.

3) Robo-ABC demonstrates a new approach to utilize and expand the potential of foundation models, while naturally enjoying the benefit of the improving ability of such models, both in retrieving objects and capturing semantic correspondence. We hope that our work will inspire future research in this direction.

2 Related Works

2.1 Learning visual affordance from human videos

Visual affordance learning aims to infer where and how to interact with diverse objects from visual inputs, bridging the computer vision and robotics fields. RGB image-based affordance learning [23, 25, 29, 43, 65, 69] methods focus on inferring affordance from images depicting human-object interactions. Another series of works focus on predicting affordance from 3D point cloud [7, 18, 32, 41, 42, 46, 57, 59, 60, 68], specifically targeting articulated or deformable objects manipulation. While taking the RGB image as sensory input, our work focuses on extracting affordance from egocentric human videos [9–11, 21], which can capture

the temporal context and motion information of human-object interaction. These allow us to better understand complex actions and generalize to new objects and scenarios. Based on the diverse reservoir of human videos, previous works have investigated to learn from them the visual representation [4, 45], grasp prior [1, 3, 27, 38, 40], and dexterous grasping skills [37, 59].

The most relevant studies to our goal are [4, 20, 34, 35, 44]. These works are dedicated to identifying the contact region with objects from human videos. However, these end-to-end approaches are subject to in-domain object categories and viewpoints. Compared to previous works, Robo-ABC extracts a small-scale memory of object interaction experiences from human videos. It allows the robot to face completely new scenes with transferable manipulation knowledge. What distinguishes our method from the previous ones is our focus on capturing the semantic correspondence beyond the seen object categories to help guide robots more accurately in complex zero-shot manipulation tasks.

2.2 Semantic correspondence for robotics

In the robotics field, previous works [12, 22, 26, 28, 58, 62] have explored leveraging semantic correspondences for robot manipulation. However, these works are somewhat limited to generalization within different instances of the same category, manual annotation, additional training, or rely on user-provided target images to perform the transfer. In this work, our goal is to achieve zero-shot generalization across object categories. Recently, foundational models such as DINO-ViT [2] and diffusion models [24, 54, 66] have demonstrated remarkable capabilities in finding semantic correspondences across objects. Specifically, features extracted from diffusion models are more versatile in mapping similar points across categories. As foundation models contain knowledge valuable for robotics tasks [15,64], we explore leveraging the semantic correspondence ability embedded within these models, eliminating the need for additional training or part segmentation. Compared to previous methods, our approach is off-the-shelf and achieves significant improvements. This provides better visual guidance for robot manipulation tasks, allowing robots to flexibly understand and infer the affordances of different categories of objects in the open world.

2.3 Generalizable robot manipulation

In the development of general-purpose robots, having generalizable manipulation capabilities is crucial, especially when applying these abilities across object categories. Facing this challenge, some works [16,17,33] focus on using point clouds as inputs, recognizing and manipulating actionable parts to achieve cross-category object manipulation. Recent works on general agents [5, 50, 63] and dexterous grasping [30, 31, 55, 61] continue to explore new possibilities. However, these methods typically require a large amount of annotated data or rely on effective part segmentation of the object. Another line of research [49, 56] utilizes foundation models and NeRF for generalizable manipulation. However, the manipulation skills of these agents are confined to a set of known instances, and

their ability to generalize falls short when encountering novel object instances. By contrast, our key motivation is to harness foundation models to extract semantic information and achieve affordance generalization beyond categories.

3 Method

In this study, we aim to obtain contact points for robotic manipulation, with a focus on generalizing to unseen objects and categories. The structure of this section unfolds as follows: Section 3.1 elucidates the process of extracting affordance knowledge from human videos, Section 3.2 discusses the retrieval of similar objects from the extracted interaction experience memory for new objects, Section 3.3 describes the process of utilizing the capabilities of the diffusion model to generalize the affordance of objects across object instances and categories, while Section 3.4 describes the process of applying the obtained affordance guidance to downstream robotic manipulation tasks.

3.1 Affordance collection from human videos

Affordance representation. We define the affordance of objects as the contact points between the human hand and the object. Humans naturally contact a variety of objects at specific points during the interaction. For example, when opening a door, the hand contacts the door handle, and such action bears fruitful information on affordance. We aim to pinpoint ***when*** and ***where*** the contact takes place given a dataset of videos. We utilize an off-the-shelf hand-object interaction detector [52] to obtain the contact state information during human-object interaction in each segment of the interaction videos.

Data Sources and Volumes. We list the sources and volumes of the affordance data used by baselines and Robo-ABC in Table 1. Since only several frames of each video are needed to extract the memory, we compute the volumes in terms of samples in alignment with other baselines. Consider a video consisting of

Table 1: The data sources and volumes comparison.

Method	HAP	HOI	HOTSPOTS	LOCATE	VRB	**Ours (Robo-ABC)**
Data Sources	EK100	EK55, EK100, EG	OPRA, EK55	AGD20K	EK100	EK100
Training Samples	18K	35K	20K	20K	54K	**864 (No training, just for memory)**

multiple frames of a person cutting vegetables: $V = \{F_1, ..., F_N\}$. We leverage a hand-object detector not only to output the pixel-level bounding boxes (bbox) of the hand B_h and the object B_o, but also to output a contact variable. This variable serves to pinpoint the exact frames where contact between the hand and the object occurs. Upon identifying the first frame F_j where the hand grasps the

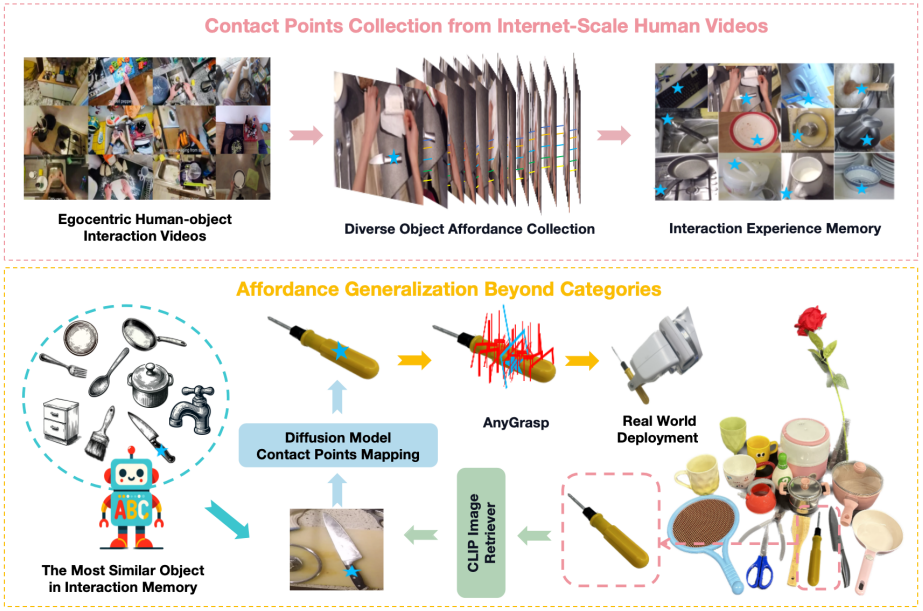


Fig. 2: Our pipeline. The top part is the process of extracting knowledge about object affordance from human-object videos. Subsequently, we store these information as interaction memory to serve as the robot's interaction experience. When facing new objects, we retrieve the most similar object from the interaction memory based on visual and semantic similarity. After obtaining the contact point information, we leverage the powerful semantic correspondence capability in the diffusion model to achieve cross-object and out-of-category affordance generalization. Finally, we select the grasp pose from all the possible poses which are generated by AnyGrasp [13] to deploy on real robots. (★ represents the positions for interacting with the object, ⊔ represents all possible grasp poses generated by AnyGrasp, ⊔ represents the grasp pose selected by ★)

knife, we use skin segmentation [51] to precisely locate the intersection area within the hand bbox B_h and objects bbox B_o. We then randomly sample from this area to obtain the contact points $P = \{p_1, p_2, ..., p_m\}$, where m is the sample number of contact points in a frame.

Contact points mapping. When collecting affordance information, we want the image to be clear, preferably without humans occluding the object. However, we need the frame of hand-object contact to deduce the contact point. To address this dilemma, we aim to map these contact points P back to a frame F_c when the object is not obscured. This can be achieved through calculating the homography matrix $\mathcal{H}_t$ between two consecutive frames to map the contact points P across different frames.

There exist several criteria for the selection of F_c: Since these frames are extracted from a video, preventing motion blur in these frames are cruial. It is also preferable to retrieve the frames near the contact frame F_j, so we can reduce the mapping error. With these considerations in mind, we set a window W around the contact frame F_j to select the frame F_c where the object's view is intact. For motion blur detection, we utilize the Laplacian operator to compute the clearest frame within the window W. We then calculate the homography matrix to map the contact point to the unobstructed frame. Lastly, we use the bounding box of the object B_o output by the hand-object detector [52] to get rid of the irrelevant surroundings of the object. We collect these cropped object images I_o and contact points P to store in the affordance memory, serving as the robot's knowledge bank of interaction experiences.

3.2 The most similar object retrieval from memory

When facing a new object, we need to retrieve similar objects from the collected memory. After capturing an image I_t with the camera, we use langsam [39] to crop the object we want to interact with from the image. An image encoder Φ is used to map the cropped object image to a feature vector z_{crop}. We also map each object image in the memory to a feature vector z_{mem} using the same encoder Φ. We use the cosine similarity of these two feature vectors as the proximity metric for retrieval.

We divide these objects into two types: The first type includes objects within the same category that have been previously encountered in the collected interaction memory, and the second type includes objects that are unfamiliar or have never been seen before. For the seen objects, we retrieve the most similar objects in the same category. For completely new object categories, we retrieve the most similar object among all objects in the memory, regardless of the category. In practice, we find that the CLIP [48] encoder suitably meets our needs, and the effects of other encoders have been tested and demonstrated in Section 5.

3.3 Semantic correspondence mapping for affordance generalization

After retrieving the most similar object from the memory and its contact points, we utilize semantic correspondence mapping to transfer such knowledge to the current scene and object. In this work, we utilize the emergent correspondence ability from the diffusion model to map the retrieved contact points to the new object, thereby guiding robotic manipulation in unfamiliar environments. More specifically, given a source image I_s, a target image I_t, and a source point p_s, we aim to find the corresponding point p_t in the target image. We follow the steps described in [53] to extract the diffusion features (DIFT) of the source image I_s and the target image I_t. The diffusion features are generated by first adding noise to the target image, then denoising it through the diffusion process, while extracting the intermediate hidden features from the U-Net simultaneously. Since the diffusion features correspond to each pixel in the target image, we can find the pixel with the highest similarity to the source point p_s in the diffusion

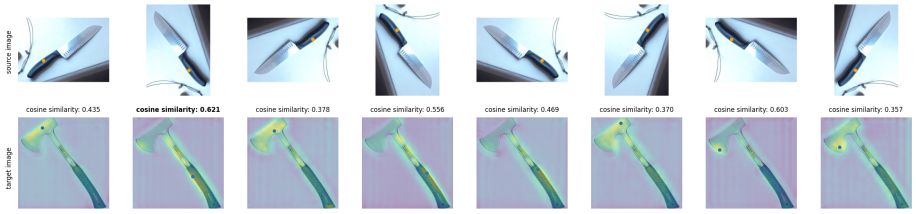


Fig. 3: Visualization of the affordance results for different orientation. The inferred correspondence point is influenced by the relative orientation between the source and target images, and those image pairs with similar orientation often have a higher similarity score and yield the correct correspondence.

features of the target image I_t. This allows us to map the contact points of the retrieved object to the current object with high fidelity and can even break the barrier caused by categories, viewpoints, and background appearance.

Specifically, we find that diffusion features are relatively prone to orientation mismatching between the source and the target image, so we deploy 8 rotation and flipping transformations to the source image, and select the result with the highest similarity in all transformed images. This simple yet effective method yields an accuracy gain of **5.4%** in our experiments. As discussed in Section 5, we found that the diffusion feature is the most effective for our task among other semantic correspondence approaches.

3.4 Deployment in the real world

After obtaining the contact points of the current object, we deploy a real robot to complete the manipulation tasks. We utilize AnyGrasp [13] to provide the grasping prior for the robot. AnyGrasp [13] can take the point cloud of the scene and provide a set of 7-DOF grasp candidates for grasping. We utilize the contact points obtained from the previous step to select the nearest grasp candidate as the final grasp pose, and use it to guide the robot to complete the manipulation task. We will discuss the details in the supplementary material.

4 Experiments

In this section, we present a comprehensive evaluation of Robo-ABC, and specifically address the following questions: 1) How does the zero-shot affordance generalization ability of Robo-ABC compare to existing end-to-end approaches for both familiar and novel object categories? 2) To what extent can Robo-ABC extrapolate affordances from a limited set of known categories to a broader spectrum of objects? 3) Upon implementation on the robot, how accurate is the semantic information provided by Robo-ABC for affordance-guided grasping, particularly in scenarios involving varying viewpoints, clustered environments,

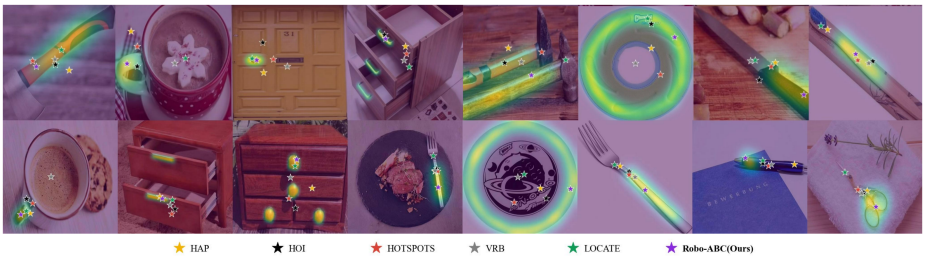


Fig. 4: Visualization of the affordance results. The highlighted areas are the ground truth masks, while stars indicate the inferred contact points of different methods. Robo-ABC(★) has the most prediction in the ground truth region.

and novel object categories? We will elaborate on these questions in the following sections. By conducting these experiments, we hope to provide a comprehensive evaluation of our method's affordance generalization ability and shed light on potential areas for further research and improvement.

4.1 Zero-Shot Affordance Generalization

Evaluation Dataset In this experiment, we select all the objects that are feasible for robotic manipulation from the "hold" action category of the AGD20K [29] dataset, ruling out objects that are too large for the robot to grasp such as beds and chairs. In this work, we focus on modeling the contact regions, and regions under the action of "hold" are more common and inclusive. Considering the prevalence of doors and drawers in everyday life, we supplemented the evaluation set with these two additional categories following the same labeling procedure. All object categories within the evaluation data set are shown in Table 5. We use the terms **seen categories** and **unseen categories** here to refer to whether objects of the same category are present in the affordance memory.

Evaluation Metrics In choosing the evaluation metrics, we claim that for the purpose of real-world object manipulation, the model's output should be contact points, instead of areas resembling a probability distribution. Therefore, we compare the accuracy of the predicted contact points across different methods. As such, we selected three metrics for our evaluation:
Success Rate (SR): The success rate is calculated as the proportion of successful points (those falling within the ground truth (GT) masks) to all output points. Since the value of the ground truth mask ranges from 0 to 255, we set a threshold of 122 to determine whether the output is feasible while providing the SR-threshold curve in Figure 6. The success rate measures the accuracy of the output directly, making it the primary metric for affordance generalization.
Normalized Scanpath Saliency (NSS): Normalized Scanpath Saliency is a straightforward method to measure the correlation between saliency maps and fixed points, calculated by averaging the normalized saliency at points of the

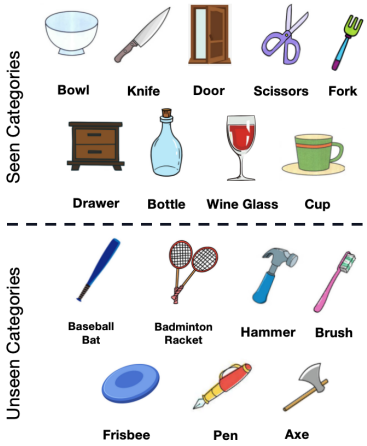


Fig. 5: Seen and unseen objects list of the evaluation dataset.

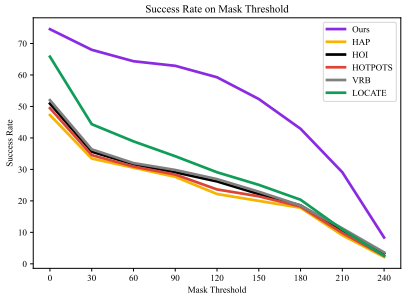


Fig. 6: Success rate on different mask threshold. We compared the success rates of Robo-ABC and baselines on different mask thresholds.Through visualization, Robo-ABC always exceeds other baselines by a large margin.

ground truth. Since we are dealing with predicted points and ground truth maps, we have modified the formula to compute the average normalized value of the ground truth map at the output points. For ground truth map $\mathcal{M}$ and output points P, the formula of NSS is as follows:

$$NSS = \frac{1}{|P|} \sum_{p \in P} \frac{\mathcal{M}(p)}{\max_{q \in \mathcal{M}} \mathcal{M}(q)} \in [0, 1] \tag{1}$$

NSS considers not only the accuracy of the output points but also their saliency. The higher the NSS, the more accurate the output points are to the center of the ground truth map.

Distance to Mask (DTM): We introduce a novel metric to compute the shortest distance between the ground truth mask region and the predicted affordance position. Using the same threshold as in success rate, we can obtain the contour $\mathcal{C}_{\mathcal{R}}$ of the ground truth mask $\mathcal{M}$. If the output point P is outside the mask area, we calculate the shortest distance from P to the contour $\mathcal{C}_{\mathcal{R}}$. If the output point P is inside the contour $\mathcal{C}_{\mathcal{R}}$, the distance to the mask is 0. DTM is then normalized by the length of the image's diagonal.

Baselines and results For the zero-shot affordance generalization experiments, we compare Robo-ABC with a series of end-to-end approaches, namely VRB [4], HOI [35], HOTSPOTS [44], and HAP [20]. Additionally, we also evaluate LOCATE [29], the work that proposed the AGD20K dataset on our benchmark. A brief introduction to these methods can be found in the supplementary material.

As we stated earlier, what we want to compare is the accuracy of the contact points rather than probabilistic distributions. Thus, we have constrained all baselines to predict points for a fair comparison. For models that output

heatmaps, we select points with the top 500 probability values, normalize their probability distributions, and sample five points from each for analysis.

As shown in Table 2, Robo-ABC achieves high success rate of 60.7%, **28.7% higher** than the second-best method, LOCATE. This demonstrates the effectiveness of our method in generalizing affordance to unseen objects. We also observe that the NSS and DTM of Robo-ABC are significantly better than other methods, indicating that the results are closer to the center of the ground truth mask.

Table 2: Main results on affordance prediction. (The results shown are all SR of different methods.) Robo-ABC surpasses all baselines by a large margin on all three metrics

Metrics	HAP [20]	HOI [35]	HOTSPOTS [44]	LOCATE [29]	VRB [4]	**Robo-ABC (Ours)**
NSS	0.231	0.239	0.236	0.283	0.242	**0.516**
SR	22.2	26.5	24.2	32.0	26.5	**60.7**
DTM	0.121	0.112	0.118	0.107	0.103	**0.045**

4.2 Cross-Category Affordance Generalization

In this experiment, we aim to showcase our method's ability to generalize the affordance of a small group of seen objects to various objects beyond its category. To this end, we fix a category of source images and provide the contact points derived from human videos. For each object of the other category, we use the same semantic correspondence setting of Robo-ABC, then obtain the target affordance, as shown in Figure 7. The similarities between the source category to the target are gradually decreasing from left to right, demonstrating the increasing challenge of affordance generalization across objects. The correspondence model of Robo-ABC can generalize even under the most challenging circumstances, like mapping the handle of a fork to the stem of a flower, which infers the great potential of our method in generalization across object categories.

4.3 Real-World Robot Experiment

Lastly, we deploy Robo-ABC along with the VRB baseline in real-world scenarios. Our method is adept at generating semantically-informed contact points. Combined with the end-to-end grasping backbone AnyGrasp, we can conduct experiments of grasping with a variety of categories.

AnyGrasp [13] is trained on a large number of real-world grasping scenarios, enabling it to generate robust and reliable grasp proposals. It takes as input a point cloud from the depth camera and outputs a set of 7-DOF grasp poses. Each detected grasp pose G is represented as $g = [\mathbf{R} \in \mathbb{R}^{3\times3}, \mathbf{t} \in \mathbb{R}^{3\times1}, w \in \mathbb{R}]$, where $\mathbf{R}, \mathbf{t}, w$ signify the rotation, translation, and width of the gripper, respectively.

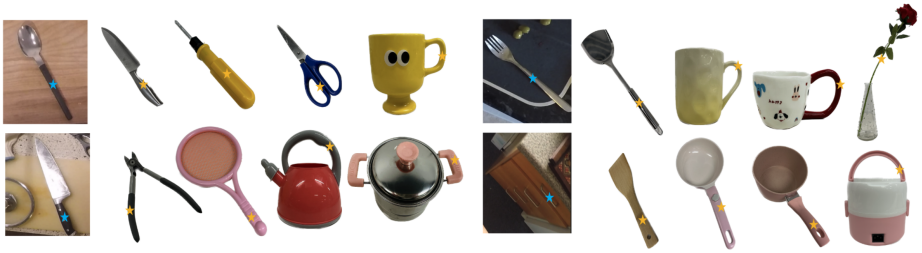


Fig. 7: Affordance generalization beyond categories visualization results. In each group of figures from left to right, the distinction of object categories gradually increases. ★ represents the contact points extracted from human videos, while ★ represents the inferred points found by Robo-ABC across object categories.

Our robotic setup consists of a Franka arm equipped with a parallel jaw gripper, and a RGBD camera mounted to provide a monocular point cloud of the scene.

Initially, we compute the contact point for the given object and ascertain its three-dimensional spatial coordinates $p^* = (x, y, z)$. The scene's point cloud is then fed into AnyGrasp to generate a set of grasp candidates G:$\{g_1, g_2, ..., g_N\}$. Among these, the pose g^*, exhibiting the minimal translational distance from point p^*, is selected as the execution pose for the robot's end-effector.

$$g^* = \arg\min_{g \in G} ||\mathbf{t}(g) - p^*|| \tag{2}$$

Our experiments are conducted over seven object categories. Robo-ABC achieves a success rate of 85.7%, compared to the baseline of 68.6%. Please refer to the supplementary material for more details.

5 Ablation Study

In this section, we examine several design choices of Robo-ABC, validating their influences by controlling variables. The selected choices include the number of seen object categories for the affordance memory, the retriever encoder, the number of retrieved images, and the selection of semantic correspondence models.

5.1 Retriever & Semantic correspondence model

For the retriever, we compare the performance of four different encoders: CLIP-B32, CLIP-B32+LPIPS, CLIP-B16, and ResNet50. The three encoders with the prefix CLIP are all based on the CLIP [47] model with different parameter sizes, and the last one is based on the ResNet50 model. Specifically, the CLIP-B32+LPIPS retriever uses the CLIP-B32 encoder with the negative LPIPS [67] loss added to the CLIP similarity, aimming to fetch the most visually and semantically similar images from the memory. For the semantic correspondence

Table 3: The performance of different retrieval methods under various correspondence models. DIFT performs the best, while SD-DINO provides comparable but less capable performance in most cases. (The results shown are all SR of different methods.)

Correspondence Methods	Retriever Encoder											
	CLIP-B32			CLIP-B32+LPIPS			CLIP-B16			ResNet50		
	DTM	NSS	SR	DTM	NSS	SR	DTM	NSS	SR	DTM	NSS	SR
DIFT [53]	**0.045**	0.516	**60.7**	**0.067**	**0.438**	50.2	**0.052**	**0.522**	**60.0**	**0.058**	**0.495**	**56.7**
SD-DINO [66]	0.052	**0.524**	58.2	0.080	0.387	**54.2**	0.077	0.418	55.6	0.073	0.360	50.5
DINO-VIT [6]	0.158	0.247	36.4	0.153	0.255	37.1	0.138	0.296	41.8	0.160	0.209	28.7
LDM-SC [24]	0.087	0.390	52.0	0.106	0.385	50.9	0.096	0.408	53.5	0.141	0.304	40.7

model, we compare the performance of four different methods: DIFT [53], SD-DINO [66], LDM-SC [24], and DINO-VIT [6]. While the first three correspondence methods are based on the diffusion models, the last one is based on the VIT. We refer the readers to the supplementary material for more details. The results are shown in Table 3 with all the metrics, and we can see that the CLIP-B32 encoder achieves the best performance, and the CLIP-B32+LPIPS encoder is slightly weaker. This indicates that the CLIP-B32 encoder alone can already provide decent results. For semantic correspondence methods, the DIFT [53] model achieves the best performance, and the DINO-VIT [6] model is the weakest, While the other models are overall less capable than DIFT [53].

5.2 Memory size of seen categories

Based on the results of the previous ablation study, we select the combination of CLIP-B32 retriever with DIFT [53] for correspondence matching. In this experiment, we fix the encoder of the retriever, aiming to validate the impact of the size of the "seen categories" on different semantic correspondence models. The results are shown in Table 4.

Table 4: The impact of categories seen in memory on the semantic correspondence models in terms of the performance of different models. (The results shown are all SR of different methods.)

Correspondence Methods	Memory Categories		
	18	36	51
DIFT [53]	56.4	56.7	**60.7**
SD-DINO [66]	53.1	53.1	58.2
DINO-VIT [6]	39.3	37.5	36.4

Table 5: The performance of different number of retrieved images. (The results shown are all SR of different methods.)

Correspondence Methods	Top K		
	1	3	5
DIFT [53]	54.5	58.2	**60.7**
SD-DINO [66]	53.5	59.3	58.2
DINO-VIT [6]	37.5	32.4	36.4

The hypothesis is that having more variety in the seen categories may improve the model's ability to generalize and correctly match contact points. However, it is also possible that there is an optimal number of categories, beyond which

model performance may start to degrade because the retriever is disturbed by a large number of categories and is unable to retrieve the most relevant images.

By varying the number of seen categories while keeping other variables unchanged, we can determine the optimal number of object categories for the model to learn from, which is crucial information for improving the system's overall performance. From the results in Table 4, we can see that the performance of the more capable models (DIFT and SD-DINO) increases as the number of seen categories increases. This result indicates that the more categories the model sees, the better the performance. However, for the less capable model of DINO-VIT, the performance decreases with regard to category numbers.

5.3 Number of retrieved images

In this experiment, we aim to validate the impact of the number of retrieved top-k images from the retriever on the performance of the semantic correspondence model. The image with the highest DIFT similarity is selected for the predicted contact point. The results are shown in Table 5. We can see that the performance of the three models increases as the number of retrieved images increases, the performance may further improve should we retrieve more images, but this is at the cost of longer inference time.

6 Conclusion

Our work focuses on enabling robotic manipulation to generalize beyond object categories, which is crucial for embodied intelligence toward the open world. We tackle the challenge of learning to interact with various objects and transferring knowledge across different categories. Inspired by the cognition process of humans, we extract an "affordance memory" containing diverse affordance information from human videos, then retrieve relevant objects from the memory based on visual and semantic similarity. Combined with a powerful diffusion model-based semantic correspondence mapping, our approach achieves significant generalization ability using only a small-scale memory information. Notably, our method achieves unsupervised zero-shot generalization without manual annotation, additional training, or human prior. We hope that our work will inspire future research in this direction and contribute to the development of embodied intelligence in the open world.

7 Limitations

In the process of extracting object affordances from human videos, we need to visualize all outputs to check the usability of the affordances produced. Although our method does not require additional training time compared to the baseline, the inference speed needs improvement as inferring semantic correspondence for each rotation takes 3.41 seconds, which is still some distance away from achieving real-time grasping.

Acknowledgments. This work is supported by the National Key R&D Program of China (2022ZD0161700).

References

1. Ananye Agarwal, Shagun Uppal, Kenneth Shaw, and Deepak Pathak. Dexterous functional grasping. In *7th Annual Conference on Robot Learning*, 2023.
2. Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.
3. Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *RSS*, 2022.
4. Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.
5. Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318. PMLR, 2023.
6. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
7. Kai Cheng, Ruihai Wu, Yan Shen, Chuanruo Ning, Guanqi Zhan, and Hao Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. *arXiv preprint arXiv:2309.07510*, 2023.
8. Sarah H Creem-Regehr and James N Lee. Neural representations of graspable objects: are tools special? *Cognitive Brain Research*, 22(3):457–469, 2005.
9. Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.
10. Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.
11. Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.
12. Norman Di Palo and Edward Johns. On the effectiveness of retrieval, alignment, and replay in manipulation. *IEEE Robotics and Automation Letters*, 2024.
13. Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 2023.
14. Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

15. Jialu Gao, Kaizhe Hu, Guowei Xu, and Huazhe Xu. Can pre-trained text-to-image models generate visual goals for reinforcement learning? *arXiv preprint arXiv:2307.07837*, 2023.
16. Haoran Geng, Ziming Li, Yiran Geng, Jiayi Chen, Hao Dong, and He Wang. Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2978–2988, 2023.
17. Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023.
18. Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. Rlafford: End-to-end affordance learning for robotic manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5880–5886. IEEE, 2023.
19. James J Gibson. The ecological approach to the visual perception of pictures. *Leonardo*, 11(3):227–235, 1978.
20. Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.
21. Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.
22. Denis Hadjivelichkov, Sicelukwanda Zwane, Lourdes Agapito, Marc Peter Deisenroth, and Dimitrios Kanoulas. One-shot transfer of affordance regions? affcorrs! In *Conference on Robot Learning*, pages 550–560. PMLR, 2023.
23. Mahmudul Hassan and Anuja Dharmaratne. Attribute based affordance detection from human-object interaction images. In *Image and Video Technology–PSIVT 2015 Workshops: RV 2015, GPID 2013, VG 2015, EO4AS 2015, MCBMIIA 2015, and VSWS 2015, Auckland, New Zealand, November 23-27, 2015. Revised Selected Papers 7*, pages 220–232. Springer, 2016.
24. Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *arXiv preprint arXiv:2305.15581*, 2023.
25. Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021.
26. Zhenyu Jiang, Hanwen Jiang, and Yuke Zhu. Doduo: Dense visual correspondence from unsupervised semantic-aware flow. In *arXiv preprint arXiv:2309.15110*, 2023.
27. Aditya Kannan, Kenneth Shaw, Shikhar Bahl, Pragna Mannam, and Deepak Pathak. Deft: Dexterous fine-tuning for real-world hand policies. *arXiv preprint arXiv:2310.19797*, 2023.
28. Zihang Lai, Senthil Purushwalkam, and Abhinav Gupta. The functional correspondence problem. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15772–15781, 2021.
29. Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931, 2023.

30. Hongyu Li, Snehal Dikhale, Soshi Iba, and Nawid Jamali. Vihope: Visuotactile in-hand object 6d pose estimation with shape completion. *IEEE Robotics and Automation Letters*, 2023.
31. Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8068–8074. IEEE, 2023.
32. Yu Li, Kai Cheng, Ruihai Wu, Yan Shen, Kaichen Zhou, and Hao Dong. Mobileafford: Mobile robotic manipulation through differentiable affordance learning. In *2nd Workshop on Mobile Manipulation and Embodied Intelligence at ICRA 2024*, 2024.
33. Yu Li, Xiaojie Zhang, Ruihai Wu, Zilong Zhang, Yiran Geng, Hao Dong, and Zhaofeng He. Unidoormanip: Learning universal door manipulation policy over large-scale and diverse door manipulation environments. *arXiv preprint arXiv:2403.02604*, 2024.
34. Yong-Lu Li, Hongwei Fan, Zuoyu Qiu, Yiming Dou, Liang Xu, Hao-Shu Fang, Peiyang Guo, Haisheng Su, Dongliang Wang, Wei Wu, et al. Discovering a variety of objects in spatio-temporal human-object interactions. *arXiv preprint arXiv:2211.07501*, 2022.
35. Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022.
36. Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023.
37. Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 6169–6176. IEEE, 2021.
38. Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pages 651–661. PMLR, 2022.
39. Luca Medeiros. lang-segment-anything. `https://github.com/luca-medeiros/lang-segment-anything`, 2023.
40. Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. *RSS*, 2023.
41. Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.
42. Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2o-afford: Annotation-free large-scale object-object affordance learning. In *Conference on Robot Learning*, pages 1666–1677. PMLR, 2022.
43. Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015.
44. Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019.
45. Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

46. Chuanruo Ning, Ruihai Wu, Haoran Lu, Kaichun Mo, and Hao Dong. Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. *arXiv preprint arXiv:2309.07473*, 2023.
47. 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
48. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.
49. Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Yunliang Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *7th Annual Conference on Robot Learning*, 2023.
50. Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.
51. Frerk Saxen and Ayoub Al-Hamadi. Color-based skin segmentation: An evaluation of the state of the art. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4467–4471. IEEE, 2014.
52. Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.
53. Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023.
54. Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023.
55. Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. *arXiv preprint arXiv:2304.00464*, 2023.
56. Yixuan Wang, Zhuoran Li, Mingtong Zhang, Katherine Driggs-Campbell, Jiajun Wu, Li Fei-Fei, and Yunzhu Li. d^3 fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation. *arXiv preprint arXiv:2309.16118*, 2023.
57. Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas J Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. In *European Conference on Computer Vision*, pages 90–107. Springer, 2022.
58. Ruihai Wu, Haoran Lu, Yiyan Wang, Yubo Wang, and Hao Dong. Unigarmentmanip: A unified framework for category-level garment manipulation via dense visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16340–16350, 2024.
59. Ruihai Wu, Chuanruo Ning, and Hao Dong. Learning foresightful dense visual affordance for deformable object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10947–10956, 2023.
60. Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, and Hao Dong. Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects. *arXiv preprint arXiv:2106.14440*, 2021.

61. Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4737–4746, 2023.
62. Zhengrong Xue, Zhecheng Yuan, Jiashun Wang, Xueqian Wang, Yang Gao, and Huazhe Xu. Useek: Unsupervised se (3)-equivariant 3d keypoints for generalizable manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1715–1722. IEEE, 2023.
63. Zhengrong Xue, Han Zhang, Jingwen Cheng, Zhengmao He, Yuanchen Ju, Changyi Lin, Gu Zhang, and Huazhe Xu. Arraybot: Reinforcement learning for generalizable distributed manipulation through touch. *arXiv preprint arXiv:2306.16857*, 2023.
64. Weirui Ye, Yunsheng Zhang, Mengchen Wang, Shengjie Wang, Xianfan Gu, Pieter Abbeel, and Yang Gao. Foundation reinforcement learning: towards embodied generalist agents with foundation prior assistance. *arXiv preprint arXiv:2310.02635*, 2023.
65. Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22479–22489, 2023.
66. Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arXiv:2305.15347*, 2023.
67. Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
68. Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper manipulation. *arXiv preprint arXiv:2207.01971*, 2022.
69. Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 408–424. Springer, 2014.