Learning Precise Affordances from Egocentric Videos for Robotic Manipulation

Supplementary Material

Contents

1. Dataset Details

In Fig. 1, we present examples from existing affordance datasets alongside our proposed Affordance Evaluation Dataset (AED), highlighting the necessity for a new evaluation dataset. UMD [12] is collected in a fixed lab environment with coarse annotations. RGBD-AFF [8] has very low resolution and clean background. IIT-AFF [13] includes humans and also annotates occluded object parts. AGD20K [11] is annotated with keypoints and transformed to coarse heatmaps with a Gaussian kernel. In contrast, AED contains natural images with pixel-wise annotations. The statistics regarding the number of images per object category are listed in Tab. 1.

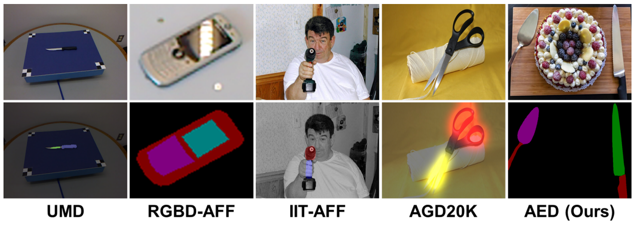
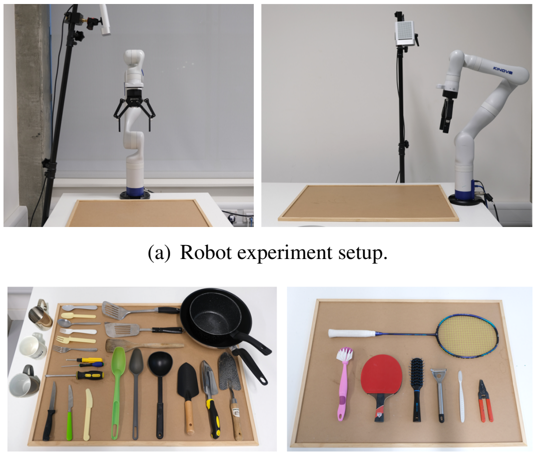


Figure 1. Examples from existing affordance datasets.

2. Implementation Details

2.1. Affordance Data Collection

We gather training data from two large-scale egocentric video datasets: Epic-kitchens [3] and Ego4d [5]. We utilize narratives to collect data of 9 object categories from Epic-kitchens, and 12 classes from Ego4d, resulting in a total of 13 object classes. For graspable point localization, correspondences between the pre-contact and contact frames are detected using the SURF descriptor [2], and the homography is then estimated by sampling at least four pairs of points with the RANSAC [4] algorithm to maximize the number of inliers. For functional point localization, the IoU threshold is set to 0.3 to detect the pre-contact frame. We set the detection thresholds to 0.1 for the hand-object detector and 0.35 for the open-vocabulary detector.



(a) Robot experiment setup.

(b) Experimental objects.

Figure 2. (a) Experimental setup. (b) Seen (left) and unseen (right) objects used in the experiments.

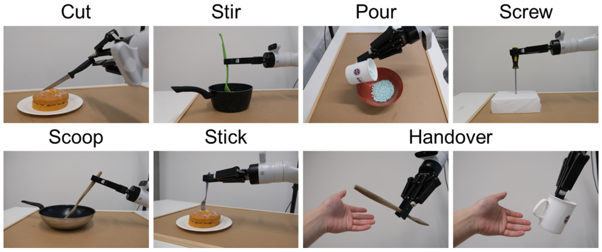


Figure 3. Illustrations of 7 tasks in the robot experiments.

2.2. Vision Experiments

All experiments are conducted on two GeForce RTX 3090 GPUs using the Adamw [10] optimizer, with a learning rate of $1e{-}3$ and batch size 8 for 15 epochs. DINOv2-base is used as the feature extractor. Collected images are first resized to 476×476 and then randomly cropped to 448×448. Both horizontal and vertical flipping are used for data augmentation. During training, LoRA is applied to all query, key, and value projection layers in the transformer block.

Total	knife	cup	scissors	hammer	fork	screwdriver	spatula	ladle	pan	shovel	spoon	drill	trowel
721	156	95	78	78	72	59	46	46	31	22	18	10	10

Table 1. Statistics on the number of images for each object on the AED.

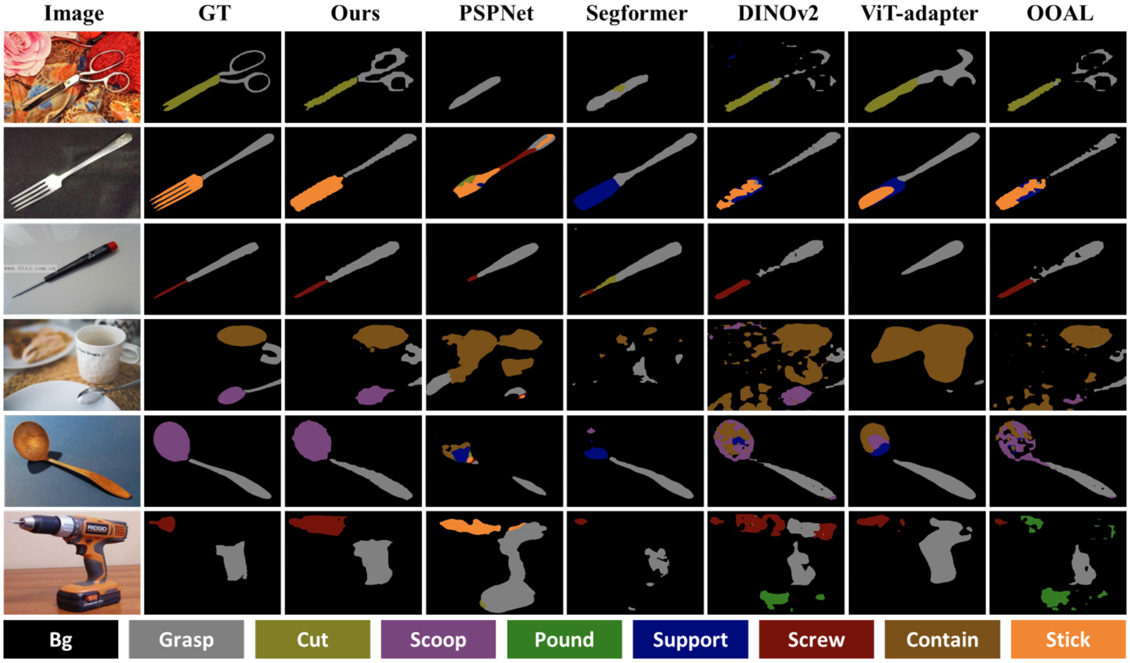


Figure 4. Qualitative comparison between our approach and other segmentation models on the AED.

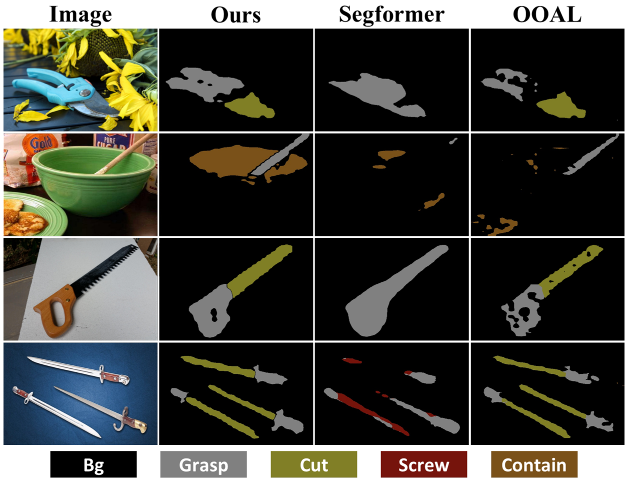


Figure 5. Qualitative comparison on unseen objects.

Focal and Dice losses are used as training objectives:

$$\mathcal{L}_{focal} = -\frac{1}{n}\sum_{i=1}^{n}\big[(1-\hat{y}_i)^{\gamma}\cdot\hat{y}_i\log(y_i) + \hat{y}_i^{\gamma}\cdot(1-\hat{y}_i)\log(1-y_i)\big], \quad (1)$$

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_i^n y_i\hat{y}_i + \epsilon}{\sum_i^n y_i + \sum_i^n \hat{y}_i + \epsilon}, \quad (2)$$

where n is the number of valid pixels in output, $\gamma = 2$ is a focusing parameter to balance easy and hard samples, and $\epsilon = 1$ is a smoothing factor that prevents division by zero and stabilizes the training. y and $\hat{y}$ represent the predicted probabilities and ground truth, respectively. Three metrics, mean intersection-over-union (mIoU), F1-score (F1), and accuracy (Acc), are adopted for evaluation.

2.3. Robot Experiments

To evaluate the effectiveness of learned visual affordances, we deploy Aff-Grasp in a 7 DoF Kinova Gen3 robot arm. The arm is equipped with a Robotiq 2F-85 parallel jaw gripper, and a calibrated Azure Kinect RGB-D camera is mounted next to the robot to capture the scene of the workspace (see Fig. 2(a)). To enable open-vocabulary affordance recognition, we utilize CLIP text embeddings as the classifier, and discard the DFI to speed up inference time. Real-world experiments are conducted with 34 diverse objects (shown in Fig. 2(b)) and 7 tasks (see Fig. 3) to evaluate

	Cut	Stir	Scoop	Screw	Pour	Stick	Handover	Total
Correct Affordance	8/9	9/9	9/9	8/9	9/9	9/9	17/18	69/72 (95.8%)
Successful Grasp	7/9	6/9	7/9	6/9	7/9	6/9	13/18	53/72 (73.6%)
Successful Interaction	7/9	6/9	5/9	5/9	6/9	5/9	11/18	46/72 (63.9%)

Table 2. Success rates for robustness evaluation.

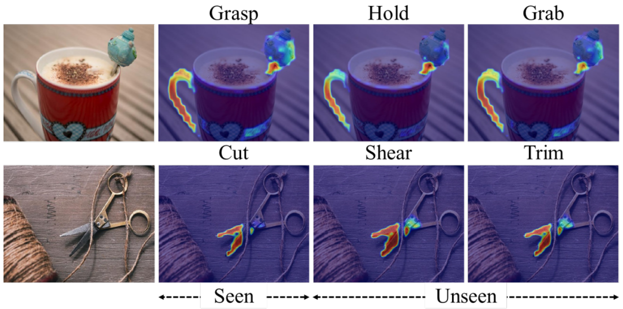


Figure 6. Qualitative examples on novel affordances.

three essential properties: accuracy, robustness, and generalization. We introduce them in detail as follows:

1. *Accuracy evaluation:* Given a single seen object on the workspace, we evaluate whether the model can recognize correct affordances of the object and perform the related affordance task. This evaluation is performed with 24 objects, each of which is repositioned 3 times during the experiment.
2. *Robustness evaluation:* Given multiple seen and unseen objects in a cluttered scene and an affordance task, we evaluate the model's ability to identify which object should be selected to perform the specific task. This requires the model to make robust predictions in the presence of distractors. The evaluation is conducted across 7 affordance tasks. Each task is tested with 3 diverse objects, except for the handover task, which is tested with 6 objects, each possessing different functional affordances. Every object is repositioned 3 times during the experiments.
3. *Generalization evaluation:* Given novel object categories not encountered during training, we evaluate if the model can still recognize the correct graspable areas. This evaluation assesses if the model can generalize the graspable affordance prediction to novel objects, which is a crucial factor in robotic manipulation. It is conducted with 7 novel objects, each repositioned 5 times.

We compare GAT with two relevant affordance grounding methods: LOCATE and Robo-ABC. LOCATE is an affordance grounding model that learns affordances from human-object interaction images using action labels as weak supervision. The method builds on DINO-ViT to identify object parts by clustering visual features from interaction regions of exocentric images, and then transfers the discovered parts to egocentric images for affordance grounding. Robo-ABC extracts object images and contact points from egocentric videos and stores these as an affordance memory. During inference, it first retrieves the most similar objects to the target and then utilizes semantic correspondence from the diffusion model to map the contact point to the current object. To ensure a fair comparison, all experiments are conducted with the Aff-Grasp framework, with only the affordance prediction component replaced.

Success rate is adopted as metric and reported from three aspects: correct affordance prediction, successful grasp, and successful interaction. For experiments in cluttered scenes, we assume that only one object is available to complete the target task. We do not perform manipulation policy learning, as it is beyond the focus of this work. Instead, we design motion primitives for each affordance and assume that the operating direction of the tool is known. For instance, in the task of "stir in the pot", the ladle is first grasped and lifted to a height of 20 cm. Next, the gripper is rotated 90 degrees along its x-axis while simultaneously moving the ladle above the pot. Finally, it lowers the ladle to a certain distance and moves in a circular trajectory around the center of the pot.

3. Experiments

3.1. Additional Results for Vision Experiments

In Fig. 4, we present additional segmentation result comparisons with all other models on the AED. In addition, we perform a qualitative comparison on images with unseen objects to explore the models' generalization ability. As displayed in Fig. 5, novel objects such as shears, saw, bowl, and sword are used. It is apparent that Segformer cannot make accurate affordance predictions for these objects. OOAL demonstrates acceptable potential on unseen objects but often produce less confident and inconsistent results. In comparison, GAT shows excellent performance on

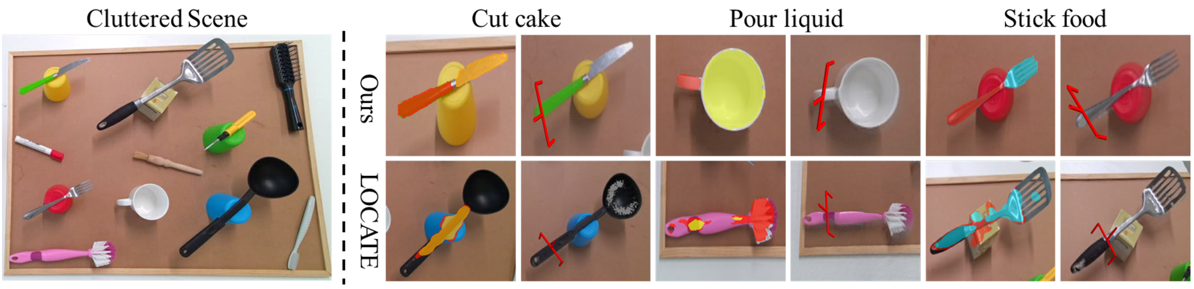
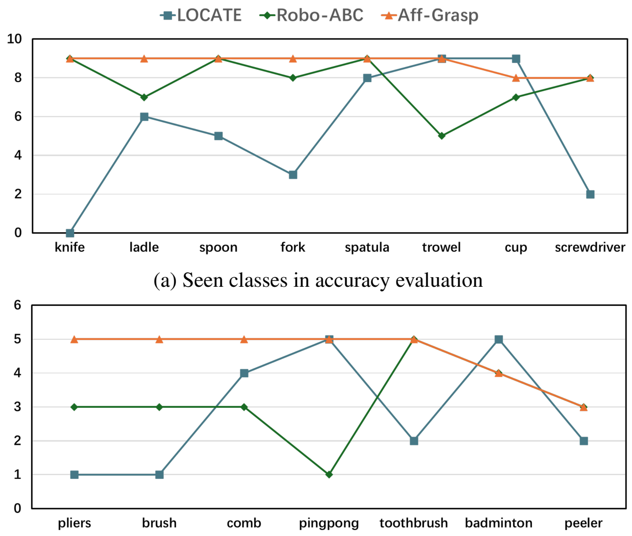


Figure 7. Qualitative comparison of affordance prediction and final grasp pose for 3D point clouds in the cluttered scene. LOCATE fails to identify related objects for desired tasks, whereas Aff-Grasp can select the correct object with accurate affordance segmentation and is not affected by cluttered scenes.



(a) Seen classes in accuracy evaluation

(b) Unseen classes in generalization evaluation

Figure 8. Success rates of correct affordance predictions for each individual object from the accuracy and generalization evaluations. The total numbers of trials are 9 and 5, respectively.

these out-of-distribution objects with much more complete segmentation maps. Furthermore, in Fig. 6, we present examples to showcase that our model can generalize to novel affordances that are synonymous with the trained actions when using CLIP text embeddings as the classifier.

3.2. Additional Results for Robot Experiments

The results for robustness evaluation is presented in Tab. 2, Aff-Grasp demonstrates strong performance in recognizing correct affordances for diverse objects, succeeding in 69 out of 72 trials. In Fig. 7, we show a visual example from the robustness evaluation, where both seen and unseen objects serve as interferences. The predicted affordance segmentation maps and corresponding grasp poses on point clouds are displayed. It is noted that LOCATE is unable to localize the correct object to execute the specified affordance task, while our model successfully identifies the matching object and predicts accurate segmentation maps. In Fig. 8, we show success rates of individual classes for accuracy and generalization evaluations. It is evident that our results are accurate and stable over all categories, while the results of LOCATE and Robo-ABC show frequent fluctuations.

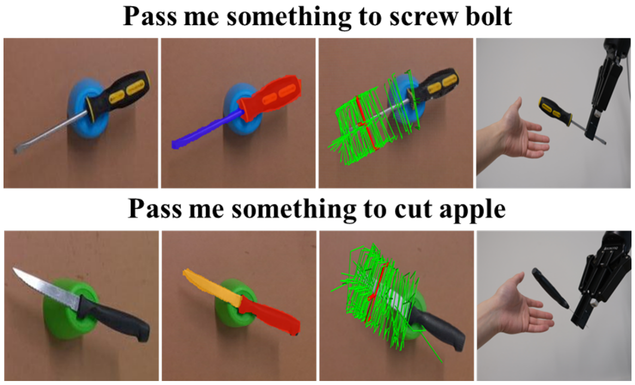

Pass me something to screw bolt
Pass me something to cut apple

Figure 9. The Aff-Grasp framework can perform the handover task by generating grasp poses within the functional parts of objects, and orienting the graspable parts towards the human hand. Green indicates all potential grasps, while red marks the final selected grasp.

Furthermore, we display affordance and grasp pose predictions for the handover task in Fig. 9. When the robot is asked to pass something to the subject for a task, Aff-Grasp generates grasp proposals based on the functional affordance mask and directs graspable parts towards the subject's hand.

3.3. Additional Ablation Studies

To further understand the effectiveness of DFI module, we perform experiments using different depth maps as input. We observe that DFI is more effective with depth representations that have low contrast. As listed in Tab. 3, the jet colormap, known for high-contrast visual effect, yields the worst results in DFI. In comparison, the less expressive

Depth map	mIoU	F1	Accuracy
Color depth (jet)	61.82	76.29	78.34
Color depth (inferno)	62.38	76.57	77.31
Color depth (viridis)	63.92	77.81	78.95
Grayscale depth	64.66	78.35	79.74

Table 3. Ablation study on different depth representations in DFI.

Embeddings	mIoU	F1	Accuracy
CLIP-B/32	66.47	79.70	79.31
CLIP-B/16	66.04	79.37	79.15
CLIP-L/14	66.91	80.02	81.09
Learnable embeds	68.62	81.09	83.51

Table 4. Ablation study on different classification embeddings: learnable or CLIP text embeddings.

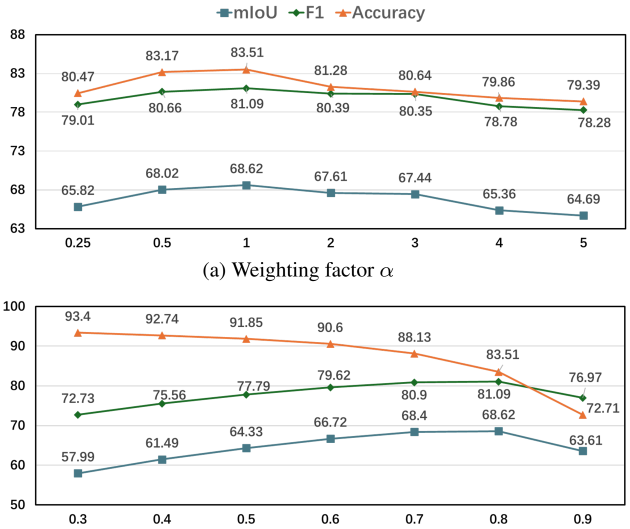
grayscale depth achieves the best performance among other colored counterparts. We speculate that grayscale input focuses more on the geometric information, whereas color depth may introduce noise to some extent.

In Tab. 4, we show the impact of different classification embeddings. The learnable embeddings yield the best results, but lose the ability to reason about unseen affordances. While performance degrades slightly when using CLIP text embeddings as classifiers, this approach retains the ability for open-vocabulary affordance segmentation. Therefore, we use learnable embeddings for vision evaluation and CLIP-L/14 text embeddings for robot experiments.

Finally, we conduct experiments with different hyperparameter settings, focusing on the threshold τ for background classification and the weighting factor α in the loss function. As presented in Fig. 10, the model obtains the highest performance in mIoU and F1-score with a weighting factor α of 1. For the background classification threshold τ, a smaller value leads to higher accuracy, as only confident predictions are counted as foreground. In this case, only mIoU and F1 score can truly reflect the performance. We thus choose 0.8 as the default threshold.

4. Limitations

Data Collection. In this work, we focus primarily on tools with distinct graspable and functional parts, which is a key stepping-stone for more general tools. However, current data collection pipeline exhibits certain limitations in handling thin and deformable objects that have small or indistinct graspable parts, such as chopsticks or wiping cloths. To address these limitations in future work, we consider integrating LLMs to acquire task priors, which will enable better distinction between graspable and functional parts. Additionally, the quality of the collected data is affected by a variety of factors. On the one hand, occlusion, motion blur, poor lighting conditions, inaccurate narrations, and unpredictable subject behavior from the video data can lead to noisy results. On the other hand, the hand-object detector and the open-vocabulary object detection model can produce incorrect predictions, further affecting the usability of the data. To mitigate these issues, we first add some constraints to reduce the error rate, such as setting high thresholds to filter out uncertain predictions. We then visualize all data samples and manually remove those with completely wrong annotations. Figure 11 displays a screenshot of the collected data. Notably, the proposed data collection pipeline is not perfect and the annotations of many samples are noisy and incomplete. Nevertheless, we retain these noisy data to assess the model's performance in this challenging situation.



(a) Weighting factor α

(b) Threshold τ for background classification

Figure 10. Ablation study on hyper-parameters.

Model Weakness. The model prediction can be susceptible to complex texture. As shown in Fig. 12, the model fails to make correct predictions when the target objects have complex textures or packaging. Also, the model sometimes confuses object parts with similar materials and shapes. For example, the head of a trowel is incorrectly recognized as having a "cutting" affordance.

Robot Experiments. Our work improves robotic grasping and interaction performance in real-world scenarios by advancing affordance prediction, as more complete and accurate object part segmentation allows the grasp estimation model to identify grasp poses with greater confidence. However, due to issues such as depth measurement errors, partial point clouds, unreliable grasp poses, and robot self-collision, correct affordance prediction does not guarantee a successful grasp, and a successful grasp does not always result in effective tool-object interaction. Since the focus

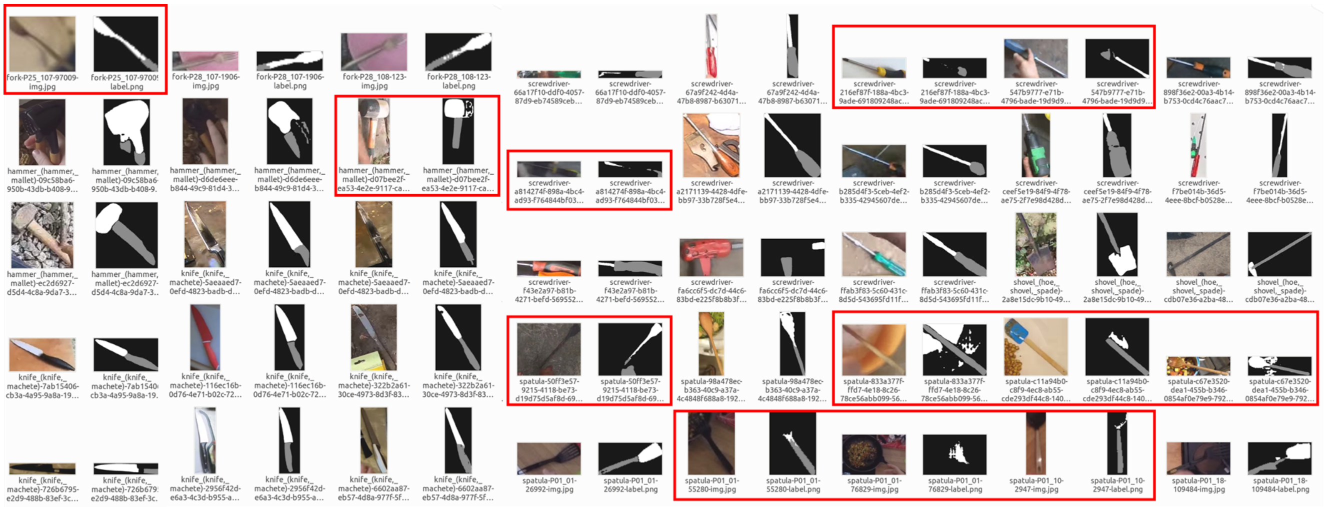

fork-P25_107-97009-img.jpg
fork-P25_107-9700 label.png
fork-P28_107-1906-img.jpg
fork-P28_107-1906-label.png
fork-P28_108-123-img.jpg
fork-P28_108-123-label.png
hammer_(hammer,_mallet)-09c58ba6-950b-43db-b408-9…
hammer_(hammer, mallet)-09c58ba6-950b-43db-b408-9.
hammer_(hammer,_mallet)-d6de6eee-b844-49c9-81d4-3…
hammer_(hammer,_mallet)-d6de6eee-b844-49c9-81d4-3…
hammer_(hammer,_mallet)-d07bee2f-ea53-4e2e-9117-ca…
hammer_(hammer,_mallet)-d07bee2f-ea53-4e2e-9117-ca…
hammer_(hammer,_mallet)-ec2d6927-d5d4-4c8a-9da7-3…
hammer_(hammer, mallet)-ec2d6927-d5d4-4c8a-9da7-3..
knife_(knife,_machete)-5aeaaed7-0efd-4823-badb-d…
knife_(knife,_machete)-5aeaaed7-0efd-4823-badb-d…
knife_(knife,_machete)-5aeaaed7-0efd-4823-badb-d…
knife_(knife,_machete)-5aeaaed7-0efd-4823-badb-d…
knife_(knife,_machete)-7ab15406-cb3a-4a95-9a8a-19…
knife_(knife,_machete)-7ab1540 cb3a-4a95-9a8a-19.
knife_(knife,_machete)-116ec16b-0d76-4e71-b02c-72…
knife_(knife,_machete)-116ec16b-0d76-4e71-b02c-72…
knife_(knife,_machete)-322b2a61-30ce-4973-8d3f-83…
knife_(knife,_machete)-322b2a61-30ce-4973-8d3f-83…
knife_(knife,_machete)-726b6795-e2d9-488b-83ef-3c…
knife_(knife,_machete)-726b679 e2d9-488b-83ef-3c.
knife_(knife,_machete)-2956f42d-e6a3-4c3d-b955-a…
knife_(knife,_machete)-2956f42d-e6a3-4c3d-b955-a…
knife_(knife,_machete)-6602aa87-eb57-4d8a-977f-5f…
knife_(knife,_machete)-6602aa87-eb57-4d8a-977f-5f…
screwdriver-66a17f10-ddf0-4057-87d9-eb74589ceb…
screwdriver-66a17f10-ddf0-4057-87d9-eb74589ceb…
screwdriver-67a9f242-4d4a-47b8-8987-b63071…
screwdriver-67a9f242-4d4a-47b8-8987-b63071…
screwdriver-216ef87f-188a-4bc3-9ade-691809248ac…
screwdriver-216ef87f-188a-4bc3-9ade-691809248ac…
screwdriver-547b9777-e71b-4796-bade-19d9d9…
screwdriver-547b9777-e71b-4796-bade-19d9d9…
screwdriver-898f36e2-00a3-4b14-b753-0cd4c76aac7…
screwdriver-898f36e2-00a3-4b14-b753-0cd4c76aac7…
screwdriver-a814274f-898a-4bc4-ad93-f764844bf03…
screwdriver-a814274f-898a-4bc4-ad93-f764844bf03…
screwdriver-a2171139-4428-4dfe-bb97-33b728f5e4…
screwdriver-a2171139-4428-4dfe-bb97-33b728f5e4…
screwdriver-b285d4f3-5ceb-4ef2-b335-42945607de…
screwdriver-b285d4f3-5ceb-4ef2-b335-42945607de…
screwdriver-ceef5e19-84f9-4f78-ae75-2f7e98d428d…
screwdriver-ceef5e19-84f9-4f78-ae75-2f7e98d428d…
screwdriver-f7be014b-36d5-4eee-8bcf-b0528e…
screwdriver-f7be014b-36d5-4eee-8bcf-b0528e…
screwdriver-f43e2a97-b81b-4271-befd-569552…
screwdriver-f43e2a97-b81b-4271-befd-569552…
screwdriver-fa6cc6f5-dc7d-44c6-83bd-e225f8b8b3f…
screwdriver-fa6cc6f5-dc7d-44c6-83bd-e225f8b8b3f…
screwdriver-ffab3f83-5c60-431c-8d5d-543695fd11f…
screwdriver-ffab3f83-5c60-431c-8d5d-543695fd11f…
shovel_(hoe,_shovel,_spade)-2a8e15dc-9b10-49…
shovel_(hoe,_shovel,_spade)-2a8e15dc-9b10-49…
shovel_(hoe,_shovel,_spade)-cdb07e36-a2ba-48…
shovel_(hoe,_shovel,_spade)-cdb07e36-a2ba-48…
spatula-50ff3e57-9215-4118-be73-d19d75d5af8d-69…
spatula-50ff3e57-9215-4118-be73-d19d75d5af8d-69…
spatula-98a478ec-b363-40c9-a37a-4c4848f688a8-192…
spatula-98a478ec-b363-40c9-a37a-4c4848f688a8-192…
spatula-833a377f-ffd7-4e18-8c26-78ce56abb099-56…
spatula-833a377f-ffd7-4e18-8c26-78ce56abb099-56…
spatula-c11a94b0-c8f9-4ec8-ab55-cde293df44c8-140…
spatula-c11a94b0-c8f9-4ec8-ab55-cde293df44c8-140…
spatula-c67e3520-dea1-455b-b346-0854af0e79e9-792…
spatula-c67e3520-dea1-455b-b346-0854af0e79e9-792…
spatula-P01_01-26992-img.jpg
spatula-P01_01-26992-label.png
spatula-P01_01-55280-img.jpg
spatula-P01_01-55280-label.png
spatula-P01_01-76829-img.jpg
spatula-P01_01-76829-label.png
spatula-P01_10-2947-img.jpg
spatula-P01_10-2947-label.png
spatula-P01_18-109484-img.jpg
spatula-P01_18-109484-label.png

Figure 11. Screenshot of the collected data. Noisy annotations are highlighted with red bounding boxes.

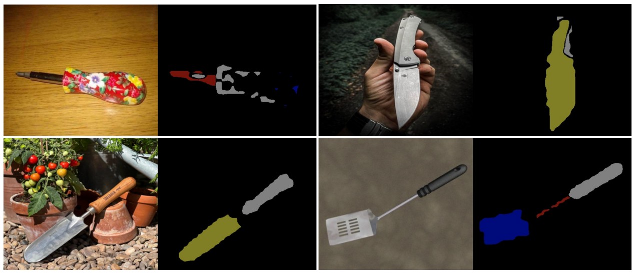


Figure 12. Failure cases. The model fails to recognize objects with complex texture and confuses parts with similar shapes and appearances.

of this paper is primarily on visual affordance learning, we did not fine-tune the grasp generation model, nor did we perform policy learning to improve the grasp and interaction success rate. Moreover, in our robot experiments, we assume the operating directions of tools are known to simplify the evaluation and design of motion primitives. To enhance practicality and scalability, a 6D object pose estimation model can be utilized to infer the operating direction.

5. Discussion

Affordance vs. Part. One may argue that parts are more direct and explicit instructions than affordances, as actions or verbs are often more abstract than semantics or nouns. A spectrum of recent work [6, 16, 18, 19] also utilize open-vocabulary part segmentation models [17, 20] and large language models [14, 15] to specify the desired grasping parts for robots. However, understanding object affordances holds great significance for embodied intelligence. Firstly, human's instructions are typically high-level and abstract. For example, we would instruct a robot to "cut an apple for me", rather than specifying "grasp the knife handle and cut the apple with the knife blade". Therefore, affordance understanding helps in the interpretation of natural instructions from humans. Second, reasoning about object parts from task instructions using large language models is time-consuming. A direct understanding of affordances can streamline the process by allowing robots to infer actionable areas from high-level instructions without extensive part-based prompting and reasoning. Thus, affordance-based approaches contribute to more intuitive and efficient interactions between robots and their environments, aligning more closely with how humans naturally communicate and perform tasks.

Points vs. Masks. In this work, we represent affordances as segmentation masks, whereas some related previous work [1, 7] represents them as points. While one may argue that using points instead of masks to acquire grasping poses is a more straightforward choice, we deem that masks are more robust and informative for the following reasons: (1) Predicting points is challenging due to their sparsity. Also, computing point correspondences is time-consuming and susceptible to variations in background and orientation. (2) A segmentation mask provides a broad region that, when combined with a grasp pose estimation model, can lead to the most confident grasp proposal. Point-based methods, on the other hand, heavily rely on the accuracy of point predictions and may fail if the predicted point is far from the object's center of mass.

Video Datasets. Although this work collects affordance data from egocentric videos, we observe that the same pipeline can also be applied to exocentric human-object interaction videos. This flexibility highlights the robustness and adaptability of our approach in different visualization perspectives. Egocentric videos provide a first-person viewpoint, which is highly beneficial for capturing the user's direct interactions with objects, allowing for a more intimate and precise understanding of affordances. On the other hand, exocentric videos, which capture interactions from a third-person perspective, can offer a comprehensive view of the context in which interactions occur. Additionally, video datasets collected in simple or laboratory environments [9, 21, 22] are preferable for ensuring high accuracy and usability of the training data. These controlled settings typically offer good lighting, background uniformity, and clear object boundaries, providing consistent and reliable data.

References

[1] 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. 6

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. 1

[3] 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. 1

[4] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1

[5] Kristen Grauman 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. 1

[6] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9488–9495. IEEE, 2024. 6

[7] Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *European Conference on Computer Vision*, pages 222–239. Springer, 2024. 6

[8] Zeyad Khalifa and Syed Afaq Ali Shah. A large scale multi-view rgbd visual affordance learning dataset. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1325–1329. IEEE, 2023. 1

[9] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 7

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[11] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022. 1

[12] 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. 1

[13] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE, 2017. 1

[14] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6

[15] OpenAI. Gpt-4v(ision) system card, 2023. 6

[16] 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. 6

[17] Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15453–15465, 2023. 6

[18] Edmond Tong, Anthony Opipari, Stanley Lewis, Zhen Zeng, and Odest Chadwicke Jenkins. Oval-prompt: Open-vocabulary affordance localization for robot manipulation through llm affordance-grounding. *arXiv preprint arXiv:2404.11000*, 2024. 6

[19] Tjeard van Oort, Dimity Miller, Will N Browne, Nicolas Marticorena, Jesse Haviland, and Niko Suenderhauf. Open-vocabulary part-based grasping. *arXiv preprint arXiv:2406.05951*, 2024. 6

[20] Meng Wei, Xiaoyu Yue, Wenwei Zhang, Shu Kong, Xihui Liu, and Jiangmiao Pang. Ov-parts: Towards open-vocabulary part segmentation. *Advances in Neural Information Processing Systems*, 36:70094–70114, 2023. 6

[21] 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*, 2022. 7

[22] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024. 7