Supplementary Material for
One-Shot Open Affordance Learning with Foundation Models

Gen Li[1] Deqing Sun[2] Laura Sevilla-Lara[1] Varun Jampani[3]

[1]University of Edinburgh [2]Google Research [3]Stability AI

A. Dataset Details

To evaluate the model's generalization ability in the challenging One-shot Open Affordance Learning (OOAL) setting, datasets with a large number of object categories are required. In addition, at least two object categories are needed for each affordance so that the model can be trained on one object and tested on the other. After an investigation of existing affordance datasets, we find only two datasets, AGD20K [1] and UMD [2], that fulfill the prerequisites and can be used to evaluate the affordance segmentation task. Specific affordance and object categories of these two datasets are shown in Tab. 1. For the unseen split, we display the object category division in Tab. 2. The model is trained on base object classes, and evaluated on novel objects categories.

Moreover, it is worth noting that annotations in AGD20K and UMD are of different types. UMD uses pixel-level dense binary maps, while the ground truth of AGD20K consist of sparse keypoints within the affordance areas, and a gaussian distribution is then applied on each point to generate dense annotation. The difference of dense and sparse affordance annotation is highlighted in Fig. 1.

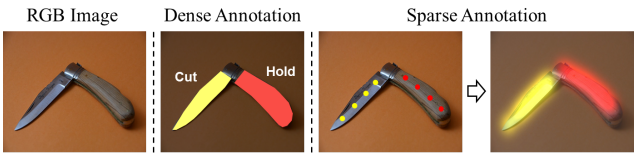

Figure 1. Different affordance annotation schemes. Dense affordance annotation is labeled as binary masks. Sparse affordance annotation is first labeled as keypoints, and then a gaussian kernel is performed over each point to produce pixel-wise ground truth.

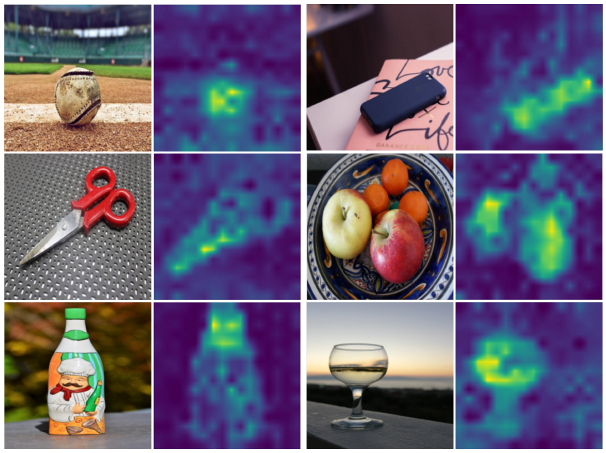

Figure 2. Visualization of CLS-guided mask.

B. Ablation Study on Hyperparameters

The proposed framework involves three primary hyperparameters, *i.e.*, the number of learnable text tokens p, vision encoder fusion layers j, decoder transformer layers t. We conduct ablation studies individually to explore the impact of these hyperparameters, as detailed in Tab. 3, Tab. 4, and Tab. 5. Notably, increasing the number of learnable text tokens up to 8 showcases a gradual improvement in performance within the seen setting, but leads to fluctuating results in the unseen setting, indicating its susceptibility to generalization when confronted with unseen objects. In terms of the fusion layers, the fusion of the last two layers demonstrates an obvious performance gain compared to the single-layer counterpart, and integrating the last three layers yields the best results. Lastly, we note that the transformer decoder can effectively improve performance in both seen and unseen setting, and a two-layer transformer decoder produces the most optimal results.

C. Additional Visualizations

C.1. Visualization of CLS-guided mask

In Fig. 2, we display the visualization of the CLS-guided mask from the proposed CLS-guided transformer decoder. It can be seen that the mask primarily concentrates on foreground objects, thus facilitating the cross-attention within salient regions.

Dataset	Affordance	Object
UMD	(7) grasp, cut, scoop, contain, pound, support, wrap-grasp	(17) bowl, cup, hammer, knife, ladle, mallet, mug, pot, saw, scissors, scoop, shears, shovel, spoon, tenderizer, trowel, turner
AGD20K	(37) beat, boxing, brush with, carry, catch, cut, cut with, drag, drink with, eat, hit, hold, jump, kick, lie on, lift, look out, open, pack, peel, pick up, pour, push, ride, sip, sit on, stick, stir, swing, take photo, talk on, text on, throw, type on, wash, write	(50) apple, axe, badminton racket, banana, baseball, baseball bat, basketball, bed, bench, bicycle, binoculars, book, bottle, bowl, broccoli, camera, carrot, cell phone, chair, couch, cup, discus, drum, fork, frisbee, golf clubs, hammer, hot dog, javelin, keyboard, knife, laptop, microwave, motorcycle, orange, oven, pen, punching bag, refrigerator, rugby ball, scissors, skateboard, skis, snowboard, soccer ball, suitcase, surfboard, tennis racket, toothbrush, wine glass

Table 1. Affordance and object classes in the UMD and AGD20K dataset. The number of classes is shown in parentheses.

Dataset	Base Objects (Train)	Novel Objects (Test)
UMD	(8) bowl, hammer, knife, mallet, mug, scissors, spoon, turner	(9) cup, ladle, pot, saw, scoop, shears, shovel, tenderizer, trowel
AGD20K	(33) apple, badminton racket, baseball, baseball bat, bench, book, bottle, bowl, carrot, cell phone, chair, couch, discus, fork, frisbee, hammer, hot dog, javelin, keyboard, microwave, motorcycle, orange, oven, punching bag, rugby ball, scissors, skateboard, snowboard, suitcase, surfboard, tennis racket, toothbrush, wine glass	(14) axe, banana, basketball, bed, bicycle, broccoli, camera, cup, golf clubs, knife, laptop, refrigerator, skis, soccer ball

Table 2. Object category division in the unseen split of UMD and AGD20K dataset. The number of categories is shown in parentheses.

p	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
2	0.774	0.568	1.710	1.119	0.457	1.434
4	0.765	0.573	1.714	1.102	**0.469**	1.449
6	0.760	0.572	1.726	1.162	0.440	1.383
8	**0.740**	0.577	**1.745**	**1.070**	0.461	**1.503**
10	0.768	**0.581**	1.726	1.111	0.460	1.463

Table 3. Ablation study on the number of learnable token p in text prompt learning.

j	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
1	1.060	0.455	1.422	1.338	0.390	1.302
2	0.748	0.576	**1.756**	1.105	0.456	1.452
3	**0.740**	0.577	1.745	**1.070**	**0.461**	**1.503**
4	0.762	**0.579**	1.713	1.129	0.453	1.401

Table 4. Ablation study on the number of fusion layers j in multi-layer feature fusion.

t	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
0	0.846	0.537	1.622	1.115	0.447	1.440
1	0.753	0.574	1.737	1.094	0.449	**1.492**
2	**0.740**	**0.577**	**1.745**	**1.067**	**0.465**	**1.492**
3	0.746	0.575	1.738	1.110	0.458	1.456

Table 5. Ablation study on the number of transformer decoder layers t.

C.2. Visualization of Unseen Affordances

In Fig. 4, we further display examples on AGD20K dataset to showcase that our model has the ability to recognize unseen affordances. It is evident that the model can consistently activate relevant affordance areas when receiving text that are previously unseen during training.

C.3. Additional Qualitative Results

In Fig. 3, we present more qualitative results on AGD20K dataset. The comparison demonstrates that predictions from our methods exhibit clear separation among object parts, while predictions from other approaches often

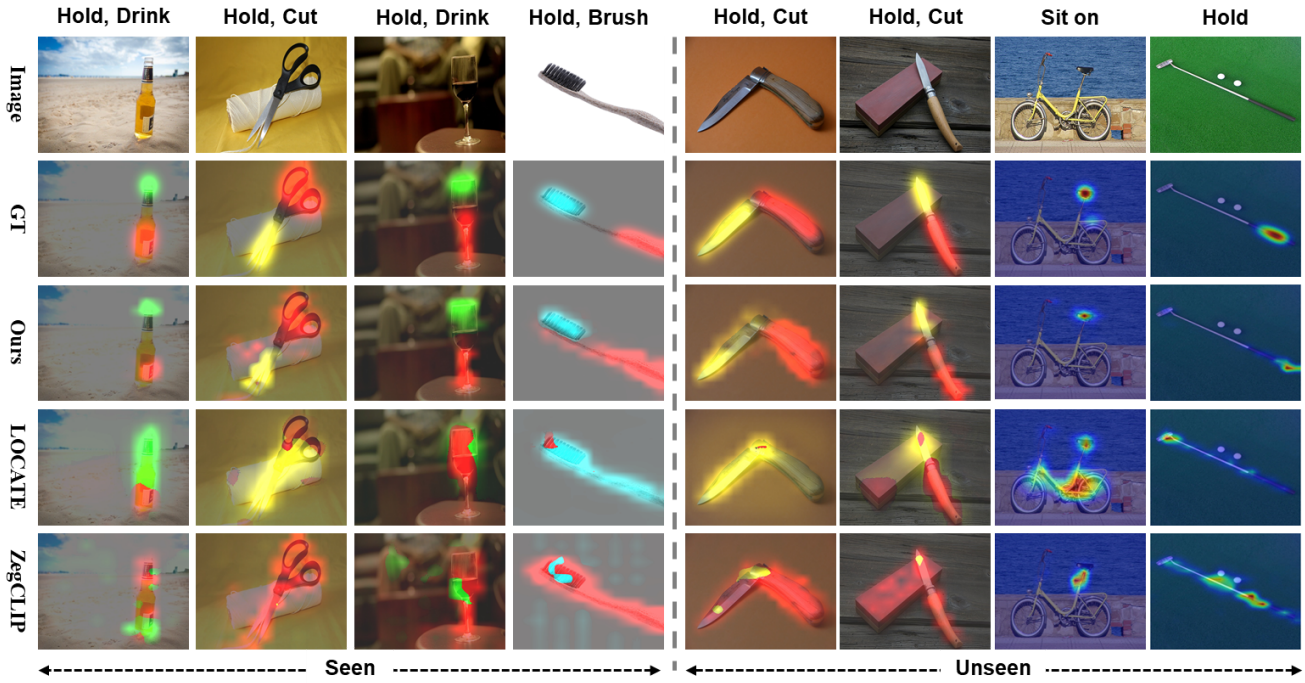

Figure 3. Additional qualitative comparison on AGD20K dataset.

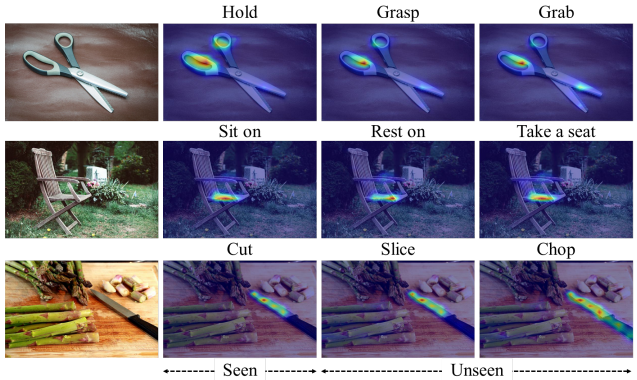

Figure 4. Qualitative examples of unseen affordance prediction on AGD20K dataset. The 2nd column shows the results on seen affordances, and the 3rd and 4th columns show results with unseen affordances.

bias towards one part or the whole object. In particular, our methods can locate very fine-grained affordance areas even for unseen objects, such as the saddle of a bicycle for "sit on", and the handle of a golf club for "hold".

D. Discussion and Limitations

This study introduces a novel problem of OOAL, and presents a framework built upon foundation models that can perform effective affordance learning with limited samples and annotations. We note that this framework can be potentially used in various applications, such as robotic manipulation and virtual reality. For instance, in robotic manipulation, the model can make reasonable affordance predictions for diverse base and novel objects, requiring minimal annotation effort. This stands in contrast to traditional methods that necessitate extensive training data or numerous simulated interaction trials to gain affordance knowledge.

Despite achieving good performance with few training samples, our framework reveals several limitations: First, while text prompt learning enhances the performance within unseen objects, it diminishes the framework's generalization capacity to unseen affordances. This occurs due to an excess of learnable tokens potentially weakening the intrinsic word similarities within the CLIP text encoder. A viable solution to this limitation involves combining the learnable prompts with manually designed prompts. Second, the performance is notably influenced by the selection of the one-shot example. Instances with heavy occlusion or inferior lighting conditions can impact the learning performance. Given the inherent challenges in learning from merely one-shot example, this limitation appears reasonable and logical.

References

[1] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. *CVPR*, 2022. 1

[2] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. *ICRA*, 2015. 1