Intermediate Connectors and Geometric Priors for Language-Guided Affordance Segmentation on Unseen Object Categories

Supplementary Material

Table 5. Dataset Statistics

Task	Train			Val			Test		
	Shape	Qst	Inst	Shape	Qst	Inst	Shape	Qst	Inst
Seen	6883	638	16120	516	58	1215	1035	174	2416
Unseen	6160	458	11558						

Table 6. **The removed object(s) for each affordance type.**

Affordance	Removed Objects	# Removed
contain	microwave, vase	508
cut	scissors	49
display	display	488
grasp	mug, scissors	199
move	table	1389
open	microwave, trashcan	297
pour	mug, trashcan	379
press	keyboard	125
stab	scissors	51
support	chair	1848
wrap_grasp	vase	381

A. Dataset Setting

To formally evaluate the generalization capability of LASO models, we adopt the unseen setting introduced in LASO [16], wherein specific affordance-object category pairs are systematically withheld from the training data. Formally, given an affordance type associated with multiple object categories, we remove selected object-affordance pairs from the training partition while preserving them in the test set. For instance, consider the affordance type "grasp", initially associated with categories such as mugs and bags. In the unseen setting, the training set excludes "grasp-mug", thus requiring the model to generalize the learned affordance knowledge from the observed category (e.g., bags) to the omitted category (e.g., mugs) during testing. Detailed definitions of the omitted affordance-category pairs are provided in Tab. 6, and comprehensive dataset statistics are presented in Tab. 5. Note that this setting explicitly assesses the model's ability to transfer affordance concepts to previously unseen object categories.

B. Detailed Per Affordance Result

We analyze GLANCE's performance across different affordance types in Tab. 7 to understand its generalization improvements in the Unseen setting. Unlike prior methods that struggle with unseen affordances, GLANCE exhibits a more consistent performance drop across Seen and Unseen settings, indicating better generalization. Affordances such as grasp, lift, and display retain relatively strong performance in the Unseen setting, likely due to their well-defined geometric priors (e.g., handles for grasp and lift). In contrast, affordances like wear and "wrap-grasp" still experience noticeable degradation, suggesting that learning deformable or implicit affordance regions remains challenging. Notably, GLANCE reduces the performance gap between Seen and Unseen affordances, particularly in affordances with high structural variation like support and push, highlighting its ability to transfer affordance knowledge beyond object categories seen during training. Moreover, even affordances not explicitly omitted in the Unseen setting show more stable performance, suggesting that GLANCE fosters a more interconnected understanding of affordance knowledge, mitigating the generalization collapse seen in prior methods. These results emphasize the model's strength in improving affordance transferability across diverse object categories.

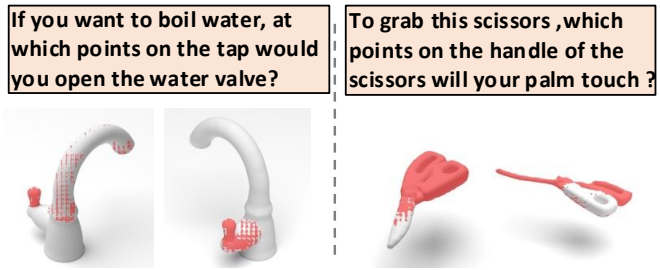

Figure 8. 2D segmentation mask provided by LISA [14].

C. Analysis of 2D Mask Results

As shown in Fig. 8 the 2D masks often contain noise, including over-segmentation (excessive coverage beyond the affordance region) and under-segmentation (missing parts of the functional area). These imperfections pose challenges for precise affordance localization, as the 2D mask serves as a critical prior for guiding segmentation in 3D space. Despite these limitations, our intra-view feature refinement helps mitigate local inconsistencies within a single view, while our cross-view consistency mechanism leverages multi-view information to refine affordance localization across different perspectives. This enables our model

Table 7. Performance of GLANCE for each affordance type.

	Metric	lay	sit	support	grasp	lift	contain	open	wrap_grasp	pour	move	display	push	pull	listen	wear	press	cut	stab
Seen	**IOU**	22.1	47.3	24.5	22.7	36.2	30.1	30.4	5.8	23.1	12.9	39.1	10.3	28.6	23.5	5.0	19.5	15.7	41.0
	AUC	88.5	97.0	91.0	83.5	93.2	89.6	92.3	69.6	90.4	79.7	93.2	85.8	84.5	93.5	69.3	94.5	94.3	100.0
	SIM	0.644	0.742	0.716	0.597	0.405	0.609	0.430	0.717	0.652	0.581	0.646	0.419	0.243	0.641	0.620	0.490	0.754	0.543
	MAE	0.101	0.067	0.085	0.114	0.066	0.095	0.050	0.130	0.085	0.134	0.075	0.078	0.039	0.111	0.146	0.046	0.073	0.022
Unseen	**IOU**	19.1	41.7	13.9	15.3	34.5	24.6	23.3	4.2	16.9	9.4	32.5	9.6	17.3	23.2	4.1	15.8	8.8	34.5
	AUC	75.3	82.2	76.6	66.3	83.0	74.8	76.1	50.4	68.9	58.7	77.0	72.0	73.9	80.6	57.6	78.8	77.6	82.1
	SIM	0.565	0.633	0.590	0.455	0.398	0.492	0.330	0.465	0.498	0.450	0.498	0.380	0.196	0.559	0.504	0.388	0.578	0.404
	MAE	0.081	0.056	0.070	0.109	0.036	0.081	0.048	0.145	0.085	0.128	0.081	0.065	0.037	0.082	0.126	0.042	0.052	0.026

to recover missing regions and suppress irrelevant areas, improving robustness even when the initial 2D masks are noisy. These findings highlight the effectiveness of our approach in reducing reliance on perfect 2D masks, making it more adaptable to real-world scenarios where segmentation errors are inevitable.

D. Limitations and Future Work

While our approach enhances generalization in language-guided affordance segmentation, several challenges remain: (1) Real-world deployment requires further validation, as sensor noise and occlusions may impact performance. (2) Multi-view rendering introduces computational overhead, though the trade-off is manageable in datasets with limited views and can be optimized through parallelization. (3) Dependence on 2D segmentation quality may lead to error propagation, although our method maintains reasonable performance even with imperfect masks. Future work could explore real-world robotic validation, lightweight alternatives to multi-view processing, and uncertainty-aware segmentation to improve robustness and efficiency.