LASO: Language-guided Affordance Segmentation on 3D Object

Supplementary Material

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

7.1. Dataset Setting

To challenge the generalization ability of the model, we create the unseen setting, by omitting samples of certain object classes from the seen setting. Specifically, for an affordance type, we omit its combination with certain objects from seen if there is more than one object class associated with this affordance type. For example, the seen partition has samples of "grasp-mug", and "grasp-bag" in its training, we create the unseen training set by removing the "grasp-mug" from the seen training set. In this case, the model is expected to learn the generalizable affordance knowledge of "grasp" and transfer it to an unseen object during testing. Note that the seen and unseen settings share the same validation and testing set. The omitted object classes for each affordance type are shown in Tab. 6:

7.2. Evaluation Metrics

In LASO, we employ four evaluation metrics: mIoU, AUC, SIM, and MAE. Here's a detailed explanation of each:

- **mIoU (Mean Intersection Over Union):** A common metric for segmentation tasks, mIoU quantifies the overlap between the predicted segmentation areas and the ground truth. It calculates the average Intersection Over Union (IoU) across all samples, reflecting the model's overall segmentation performance.
- **AUC (Area Under the Curve):** In the context of LASO, AUC assesses the model's ability to differentiate between the affordance and non-affordance parts of an object. It evaluates the model's classification performance at various thresholds, indicating how well the model can discern relevant object parts across different scenarios.
- **SIM (Similarity):** For LASO, this metric evaluates how closely the model's segmentation aligns with the actual affordance area mentioned in the question. It provides insight into the model's effectiveness in interpreting the question and accurately mapping it to the corresponding spatial regions.

$$\text{SIM}(Y, M) = \sum_{i}^{n} \min(Y_i, M_i), \tag{11}$$

$$\text{s.t.} \sum_{i}^{n} Y_i = \sum_{i}^{n} M_i = 1, \tag{12}$$

  where Y and M denote the ground truth and predictive segmentation, respectively. n is the total number of points.
- **MAE (Mean Absolute Error):** In LASO, MAE measures the average magnitude of errors in the model's affordance segmentation, without considering the error direction. It's a key indicator of the model's overall accuracy in segmenting the object part related to the query, reflecting how precisely the model can follow the affordance cues provided in the natural language question.

$$\text{MAE}(Y, M) = \sum_{i}^{n} |Y_i - M_i|, \tag{13}$$

  where n is the total number of points. Y and M are the ground truth and predictive segmentation, respectively.

7.3. Baselines

Given the absence of existing research using paired question-point cloud data for object affordance segmentation, we adapt two methods from 3D cross-modal research–3D-SPS [26] and IAGNet [44]; and two from referred image segmentation–ReferTrans [15] and RelA [22], for a comprehensive evaluation of our LASO task.

- **IAGNet [44]** Closely related to LASO, IAGNet was developed for grounding 3D object affordance from 2D image interactions. We adapted IAGNet for LASO by replacing its image backbone with a language model, while retaining the rest of its structure.
- **3D-SPS[26]** Originally a 3D visual grounding method designed to locate target objects in point cloud scenes based on language descriptions, 3D-SPS progressively selects keypoints under linguistic guidance. To fit LASO's segmentation framework, we removed its bounding box prediction module, adapting the remainder to process object point clouds for affordance segmentation.

Table 7. Performance of PointRefer 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	18.5	39.3	20.2	18.0	29.1	23.8	24.1	4.6	18.2	10.4	32.0	8.1	23.3	19.0	3.9	15.2	12.3	32.8
	AUC	87.7	96.2	90.3	82.7	92.4	88.8	91.5	68.9	89.7	78.9	92.4	85.0	83.7	92.7	68.6	93.7	93.5	99.2
	SIM	0.642	0.739	0.714	0.595	0.403	0.607	0.428	0.714	0.65	0.579	0.644	0.417	0.241	0.639	0.618	0.488	0.751	0.541
	MAE	0.101	0.067	0.085	0.115	0.066	0.096	0.05	0.131	0.086	0.134	0.076	0.078	0.039	0.111	0.146	0.046	0.073	0.022
Unseen	IOU	16.5	32.7	10.5	11.5	27.0	18.6	17.3	2.8	12.7	6.7	24.2	6.9	13.2	17.7	2.8	11.8	6.5	25.8
	AUC	74.8	81.7	76.1	65.8	82.5	74.3	75.6	49.9	68.4	58.2	76.5	71.5	73.4	80.1	57.1	78.3	77.1	81.6
	SIM	0.563	0.630	0.588	0.453	0.396	0.490	0.328	0.462	0.495	0.448	0.496	0.378	0.194	0.557	0.502	0.386	0.576	0.402
	MAE	0.081	0.056	0.070	0.109	0.036	0.082	0.048	0.145	0.085	0.128	0.081	0.065	0.037	0.082	0.126	0.042	0.052	0.026

- **ReLA [22]** Originating from image-based referring expression segmentation, RelA's goal is to segment objects as described by a language expression. For LASO, we substituted its image backbone with a 3D counterpart and adapted the image region features to grouped point features, preserving its cross-modal fusion and mask decoding strategies.
- **ReferTrans [15]** A versatile, transformer-based architecture for image-based expression segmentation, ReferTrans was adjusted for LASO by replacing the image backbone with a 3D one and modifying it to support only mask prediction, removing the detection head.

7.4. Detailed Result

To have a deeper understanding of PointRefer's learning pattern, we analyze its performance on each affordance type, and the result is presented in Tab. 7. Our observations are as follows:

- **Comparative Analysis.** The comparative analysis of PointRefer's performance across different affordance types offers critical insights into the model's learning dynamics. In both the "Seen" and "Unseen" datasets, affordance types like "lift", and "display" show strong results, especially in metrics such as IOU and AUC. This trend might be attributed to the limited variety of objects these affordances cover and their relatively straightforward patterns, like a "handle" for "lift". Conversely, affordance types like "wear," "push", and "wrap_grasp" exhibit less satisfactory performance, likely due to the absence of a distinct functional part or the need for learning deformable patterns, posing a challenge in determining the focus areas for segmentation. Moving forward, it's clear that our efforts should concentrate on enhancing the model's capability to handle these less-performing affordance types, addressing their unique challenges to achieve a more comprehensive and robust performance.
- **Trend Analysis.** Overall, PointRefer showcases superior performance in the "Seen" dataset compared to the "Unseen" across various affordance types, yet it maintains a consistent pattern in both datasets. This uniformity highlights PointRefer's capacity to generalize and effectively apply learned affordance knowledge to previously unseen objects. Upon closer inspection of the "Seen" and "Unseen" results, we find that certain affordances, like "grasp", exhibit stable performance, indicative of a well-generalized understanding. In contrast, performance variations in affordances such as "support" in the "Unseen" setting reveal inherent challenges. We believe that the broader coverage of object categories for affordances like "grasp" (encompassing seven different categories in the "Seen" setting) as opposed to "support" (covering only two categories) enables the model to develop a more generalized understanding, leading to enhanced performance in "Unseen" scenarios.
  Notably, even when we exclude objects related to specific affordances in the "Unseen" setting, there is an observed decline in performance for all affordance types, including those that remain unchanged. This indicates that the model captures a generalized affordance knowledge that impacts its performance across the board, suggesting an interconnected learning approach that influences its efficacy across all types of affordance.

7.5. More Visualization Results

In Figure Fig. 9, we present additional visualizations of PointRefer's predictions. Notably, PointRefer demonstrates the capability to yield diverse segmentation outcomes for the same type of object (e.g., bottles, bowls, knives) in response to different questions. This variability underscores the effectiveness of our question-guided approach in the model's design.

Furthermore, PointRefer effectively segments the affordance parts of objects within the same category, accommodating various shapes and sizes. For instance, when tasked with identifying the "door knob" on different doors, PointRefer consistently and accurately segments these regions, regardless of their distinct appearances. This ability to adapt to different affordance representations in similar object categories highlights PointRefer's robust generalization capabilities.

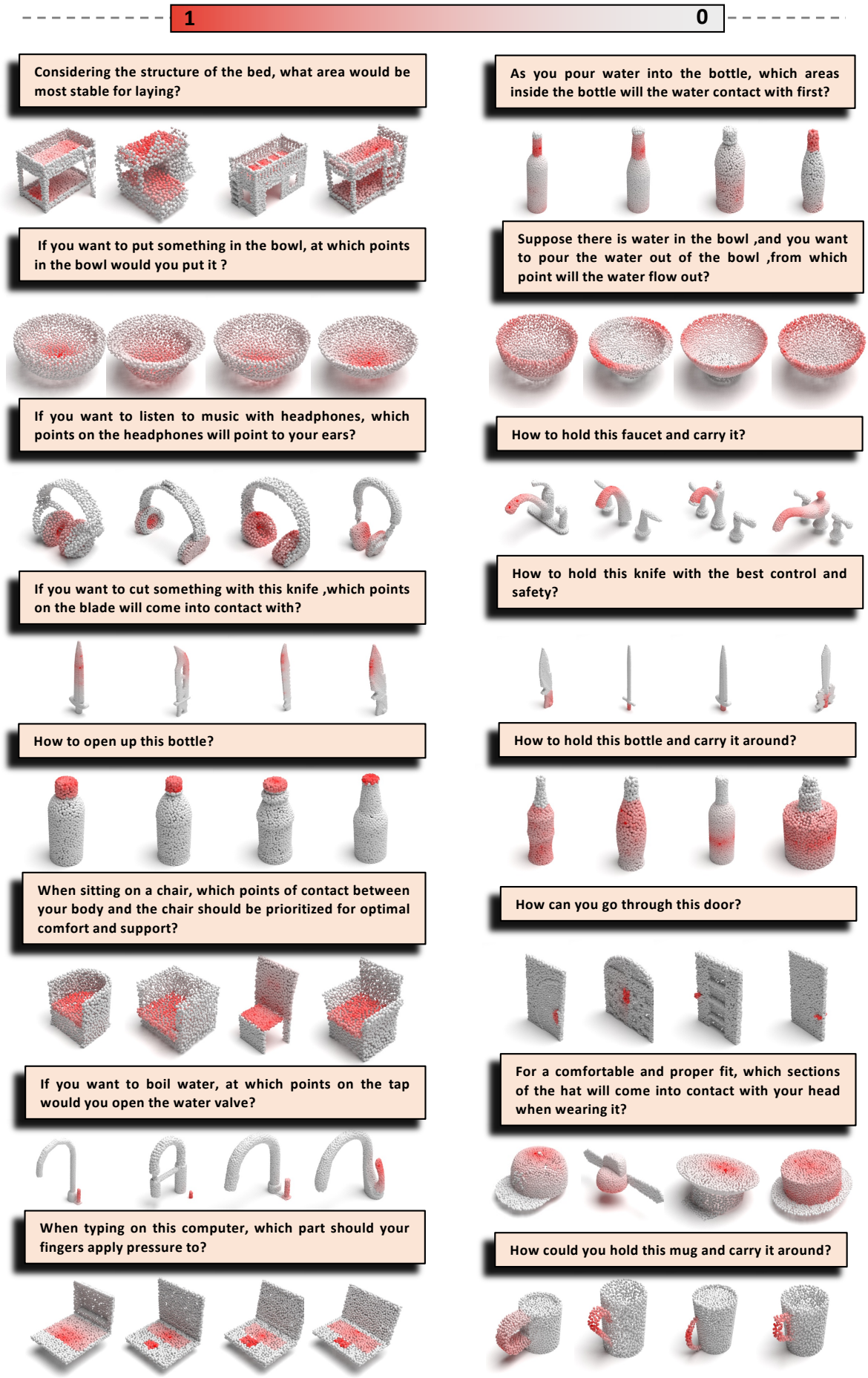

Figure 9. Case-Study of PointRefer's segmentation. Each showcase comes with one question and four shapes, showing the generalization of the prediction. The segmented affordance part is highlighted in red.