QueryMe: Query-Driven Open-Vocabulary 3D Object Affordances Grounding from Multimodal Evidence

Supplementary Material

Table 4. Comparison of performance metrics across different methods: **Ours**, **w. ROPE** with 3D ROPE position encoding, and **Baseline** with direct concatenation of text, 3D HOI, and point cloud object features. The results are reported for three settings: Seen, Unseen Object, and Unseen Affordance.

	Metric	Ours	w. ROPE	Baseline
Seen	**AUC**	**92.34**	89.42	86.59
	aIOU	**39.39**	36.79	33.06
	SIM	**0.683**	0.624	0.617
	MAE	**0.061**	0.099	0.087
Unseen-Obj	**AUC**	**83.03**	76.09	76.41
	aIOU	**21.76**	20.43	18.72
	SIM	**0.420**	0.397	0.358
	MAE	**0.118**	0.135	0.135
Unseen-Aff	**AUC**	**74.00**	68.40	66.72
	aIOU	**13.76**	12.09	10.01
	SIM	**0.316**	0.292	0.255
	MAE	**0.097**	0.141	0.150

6. More Ablation Studies

We conducted more detailed ablation experiments, as shown in Table 1. In our approach, we utilized MLP to process the position encoding of query tokens. Additionally, we explored using 3D ROPE [31] for encoding the positions of query tokens. Specifically, we partitioned the feature dimension into three roughly equal segments corresponding to the x, y, and z axes, and applied 1D ROPE [31] to each segment. However, the results did not meet expectations, particularly in the Unseen scenarios, where using 3D ROPE led to significant degradation in performance. We hypothesize that the precise position encoding provided by ROPE makes it challenging for the query tokens to generalize the learned affordance knowledge to other positions.

Furthermore, we report the performance of the baseline method, where we directly concatenated text, 3D HOI, and point cloud object features to predict affordance grounding. This approach yielded significantly lower performance compared to our proposed method.

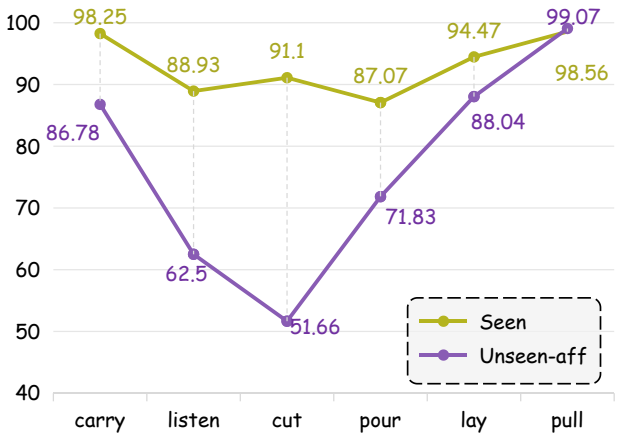

Figure 6. **Performance comparison** of affordance grounding AUC metrics between Seen and Unseen affordance settings across different categories.

7. Affordance Knowledge Differences

We compare the differences in affordance grounding AUC metrics between Seen and Unseen affordance settings for the same categories. In the Seen setting, we achieved strong performance across all categories. Notably, in the Unseen affordance setting, we observed that for affordances such as *carry*, *lay*, and *pull*, which involve larger, more planar regions, the model performed relatively well, with results approaching those in the Seen setting. In fact, for the *pull* category, the performance even exceeded that of the Seen setting. However, for the *cut* category, the results were significantly worse. This discrepancy may arise from the fact that the affordance region for *cut* is not clearly defined after 3D mapping of the HOI image, making it difficult for the model to discern relevant features. Moreover, the textual information available for unseen affordances is often limited, and the current method struggles to effectively infer the affordance regions for unseen categories. And despite using adaptive attention to mitigate noise, monocular reconstruction errors and background clutter remain significant challenges, especially in occluded or cluttered scenes. These issues can distort geometric representations, making it difficult to accurately predict affordance regions for partially obscured objects or those in complex environments.

In future work, we plan to explore methods such as continual learning to improve the model's ability to reason about finer-grained affordance regions in unseen contexts, ultimately enhancing its generalization to previously unseen affordance categories.