1. GRASPNET-RIS

To evaluate the generalization of RIS in the context of human-robot interaction, we generate referring expressions for a subset of images from GraspNet [1] using Shikra [2]. We collect the first frame for every scene from the subsets train_1, test_seen, test_similar, and test_novel. Subsequently, we select objects with clear class definitions and typical appearance, resulting in 50 categories. For instance, we exclude the *072-k_toy_airplane* which looks completely different from a conventional airplane. We generate the text expressions as follows:

- Select objects belong to chosen categories and calculate their bounding boxes based on the instance masks.

- Feed the image and each bounding box to Shikra with three different prompts, resulting in three referring expressions for each object.

- Pick the most precise and diverse description and make adjustments to ensure it accurately corresponds to the object if necessary.

In total, we generate 497 referring expressions for 105 images from GraspNet, which we denote as GraspNet-RIS.

Table 1. Zero-shot performance on GraspNet-RIS.

Train Dataset	Method	GraspNet-RIS				
		mIoU	oIoU	Prec@0.5	Prec@0.7	Prec@0.9
RefCOCO	LAVT	35.17	31.51	37.10	28.43	10.69
	Ours	**39.27**	**34.46**	**41.94**	**30.65**	**12.30**
RefCOCO+	LAVT	37.43	32.18	39.92	29.03	**13.51**
	Ours	**38.70**	**32.41**	**42.54**	**30.04**	11.29
RefCOCOg	LAVT	36.90	35.10	39.72	33.87	14.31
	Ours	**44.53**	**37.06**	**48.19**	**37.30**	**15.52**
ReferIt	LAVT	38.47	33.10	40.12	26.41	7.06
	Ours	**43.63**	**37.36**	**47.58**	**35.48**	**8.06**

Tabel 1 shows the zero-shot performance on GraspNet-RIS of models trained on four public datasets. Our method achieves clear improvements in all settings, e.g., our method notably outperforms LAVT by 5.16% mIoU and 4.26% oIoU when trained on ReferIt. It demonstrates that our method also generalizes well to scenarios with large domain shifts in practical applications.

2. REFERENCES

[1] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *CVPR*, 2020, pp. 11444–11453.

[2] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao, "Shikra: Unleashing multimodal llm's referential dialogue magic," *arXiv preprint arXiv:2306.15195*, 2023.