Language-Regularized Concept Learners

Supplementary Material

The appendix is organized as the following. In Appendix A, we specify the prompts used to query LLMs in LARC. In Appendix B, we present additional visualizations of LARC's performance. In Appendix C, we include additional results of LARC on different levels of box prediction noise. In Appendix D, we discuss the ScanRefer [4] dataset.

A. Prompts

Below, we provide the prompts used to query large language models, specifically, GPT-3.5 [3], for concepts that satisfy LARC's constraints.

Symmetry and exclusivity. We use the following prompt to categorize relational concepts, where *[relations]* is the list of relational concepts automatically extracted from the input language by LARC's semantic parser:
We define two kinds of spatial relations: Asymmetric relations are relations that don't exhibit reciprocity when the order of the objects is reversed. Symmetric relations are relations that exhibit reciprocity when the order of the objects is reversed. Here are some relations: [relations]. For each relation, specify whether it is a symmetric relation or an asymmetric relation.

Synonyms. We use the following two-round query to find visually similar synonyms in object categories, where the *[object categories]* list is automatically extracted:
First round: *Here are some object categories: [object categories]. List categories that have similar meanings.*
Second round: *Within each group, list categories that have similar appearances.*

B. Visualizations

In this section, we present additional visualizations of LARC's performance. First, we compare LARC's predictions to that of prior works on the ReferIt3D [1] dataset. Then, we provide execution trace examples of LARC. After, we demonstrate failure cases of LARC and include analyses. Finally, we show examples of VoteNet [10] object detections in comparison to ground truth bounding boxes.

Comparison to prior works We present examples of LARC's predictions as well as baselines' on the ReferIt3D [1] dataset. We see samples in Figure 1 where LARC outperforms baselines, including NS3D [7], BUTD-DETR [9], MVT [8], LAR [2], TransRefer [6], and LangRefer [11], in the naturally supervised 3D grounding setting.

Execution traces In Figure 2, we present examples of LARC's execution trace. LARC first parses input instruction utterances into symbolic programs, then hierarchically executes each modular program to retrieve the answer.

Failure cases We provide several examples of LARC's failure cases in Figure 3. In the top row, we see cases where LARC finds target objects of the correct object category, but with incorrect relations. In the bottom row, we see cases where LARC yields target objects of incorrect object categories. LARC is likely to fail in 3D visual grounding when the target object category is one without data-augumented synonyms during training, as it is difficult to learn with few examples in the naturally supervised setting.

VoteNet detections In Figure 4, we show examples of VoteNet [10] object detections, used in our low guidance setting, in comparison to ground truth bounding boxes. We see that VoteNet detections often result in incomplete point clouds, due to size corruption or center shift. This noise leads to additional challenges in 3D visual grounding; however, VoteNet detections significantly reduce the amount of labelled 3D data required during inference.

C. Experiments

Noisy detection experiments. We report results of NS3D and LARC over 6 different levels of box prediction noise in Table 1, with each column representing ratio of perturbation on the original box. LARC consistently improves NS3D under all settings.

Noise level	0.0	0.1	0.2	0.3	0.4.	0.5
NS3D	27.6	22.9	20.7	19.6	13.8	10.7
LARC (Ours)	**36.6**	**35.6**	**33.5**	**30.2**	**24.7**	**20.1**

Table 1. Comparisons under different levels of box prediction noise.

D. ScanRefer

Here, we describe how LARC uses the ScanRefer [4] data for zero-shot transfer from ReferIt3D [1].

Data construction We first create a subset of ScanRefer with queries that contain the same objects and relations as in ReferIt3D, such that we can run all method inference-only. This ScanRefer subset consists of 384 unseen utterances, on the same ScanNet [5] scenes.

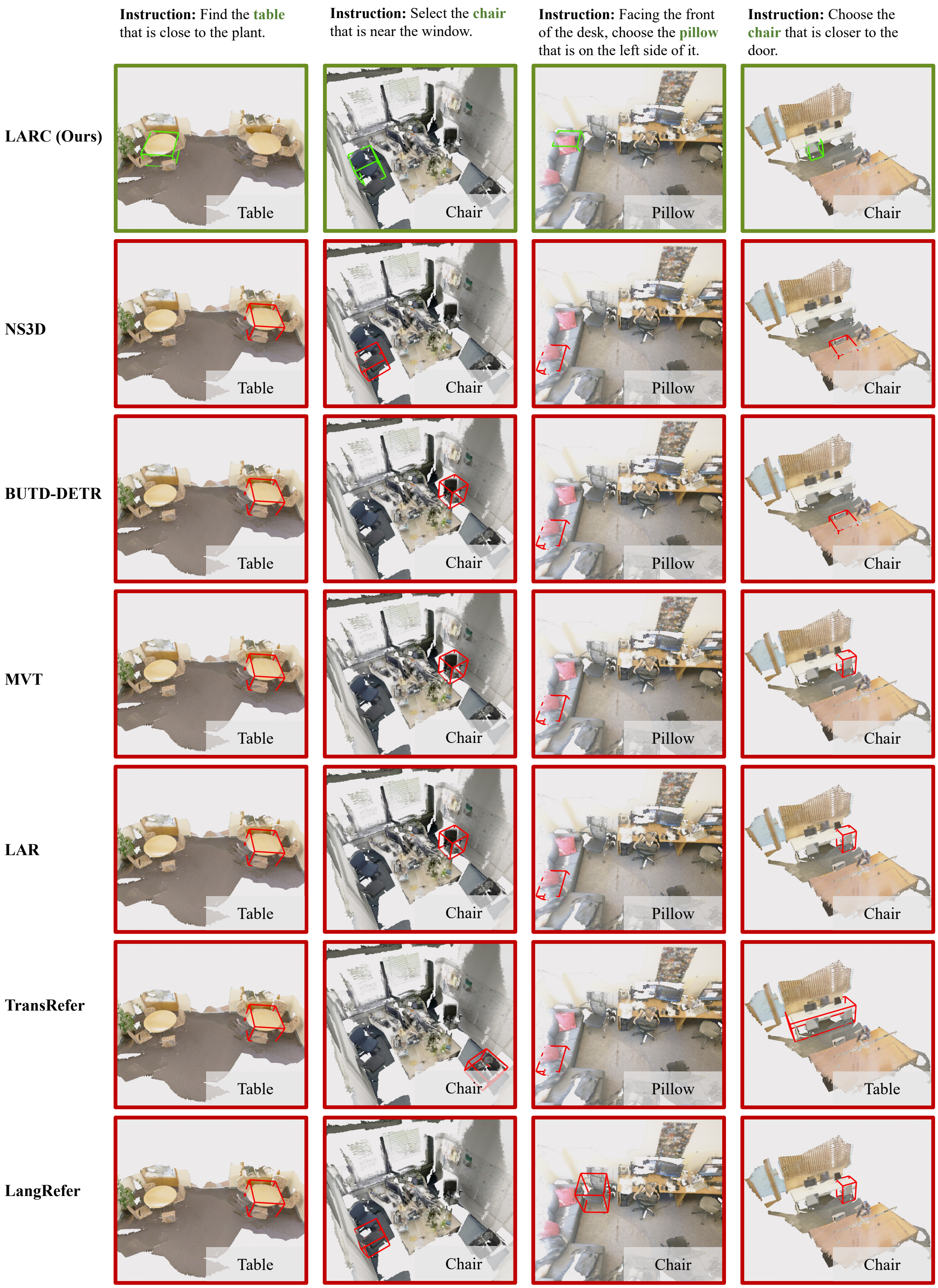
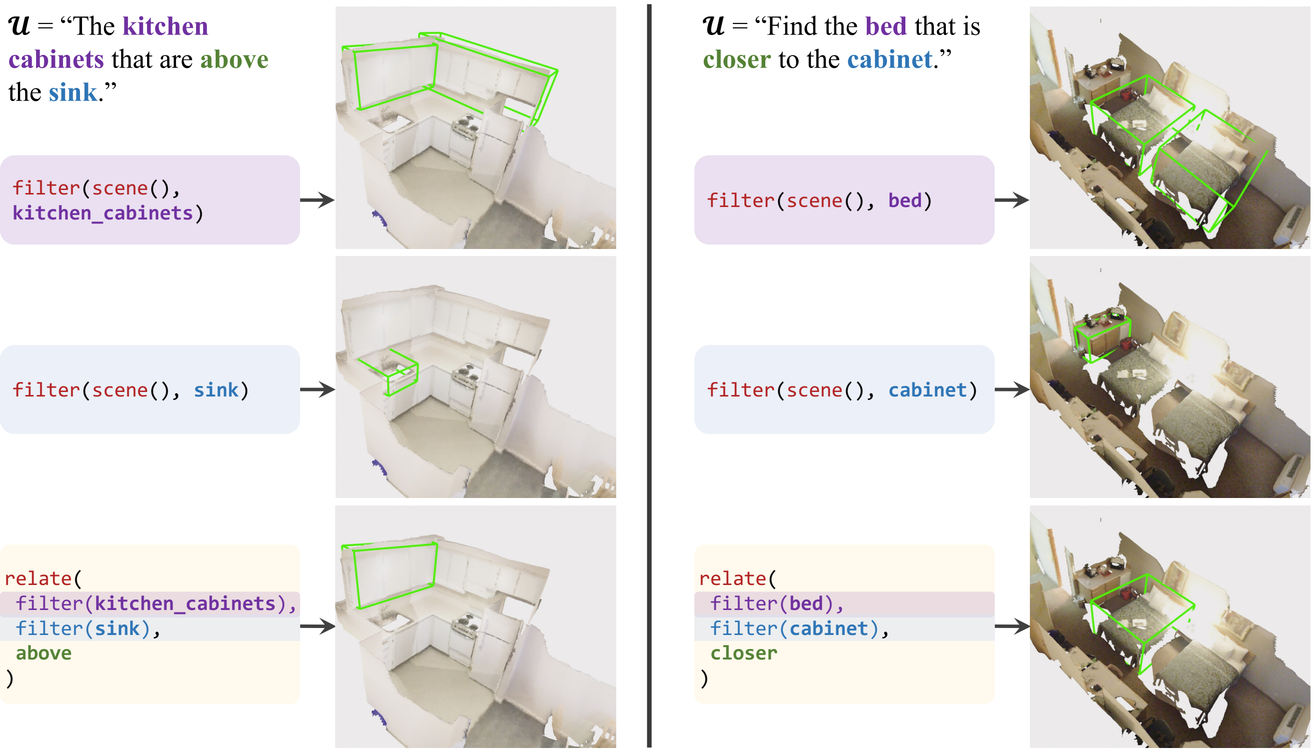

Instruction: Find the table that is close to the plant.
Instruction: Select the chair that is near the window.
Instruction: Facing the front of the desk, choose the pillow that is on the left side of it.
Instruction: Choose the chair that is closer to the door.
LARC (Ours)
Table
Chair
Pillow
Chair
NS3D
Table
Chair
Pillow
Chair
BUTD-DETR
Table
Chair
Pillow
Chair
MVT
Table
Chair
Pillow
Chair
LAR
Table
Chair
Pillow
Chair
TransRefer
Table
Chair
Pillow
Table
LangRefer
Table
Chair
Chair
Chair

Figure 1. LARC's performance compared to prior works in the naturally supervised setting; each column shows every model's prediction for a given instruction.



$\mathcal{U}$ = “The kitchen cabinets that are above the sink.”

```
filter(scene(),
kitchen_cabinets)
```

```
filter(scene(), sink)
```

```
relate(
 filter(kitchen_cabinets),
 filter(sink),
 above
)
```

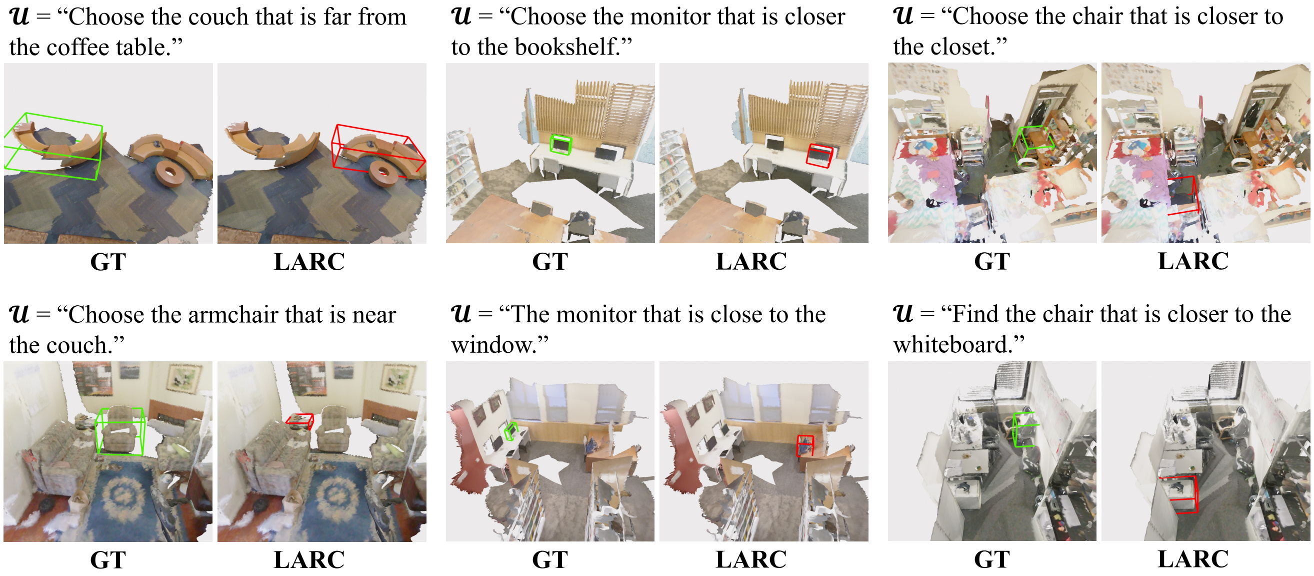
$\mathcal{U}$ = “Find the bed that is closer to the cabinet.”

```
filter(scene(), bed)
```

```
filter(scene(), cabinet)
```

```
relate(
 filter(bed),
 filter(cabinet),
 closer
)
```

Figure 2. LARC’s neuro-symbolic framework executes symbolic programs hierarchically to retrieve the target answers.



$\mathcal{U}$ = “Choose the couch that is far from the coffee table.”

$\mathcal{U}$ = “Choose the monitor that is closer to the bookshelf.”

$\mathcal{U}$ = “Choose the chair that is closer to the closet.”

$\mathcal{U}$ = “Choose the armchair that is near the couch.”

$\mathcal{U}$ = “The monitor that is close to the window.”

$\mathcal{U}$ = “Find the chair that is closer to the whiteboard.”

Figure 3. LARC can fail in understanding 3D relations (top row) or 3D object categories (bottom row); its modularity enables such analyses.

Implementation To transfer learned concepts to ScanRefer, we use GPT as LARC’s semantic parser to generate programs from input language. The programs are executed as described in the main paper. LARC relies on the generalization abilities of LLMs to zero-shot transfer to ScanRefer, by decomposing new language into learned programs, without requiring any additional training or finetuning of neural networks. In comparison, end-to-end methods significantly underperform when faced with unseen input language.

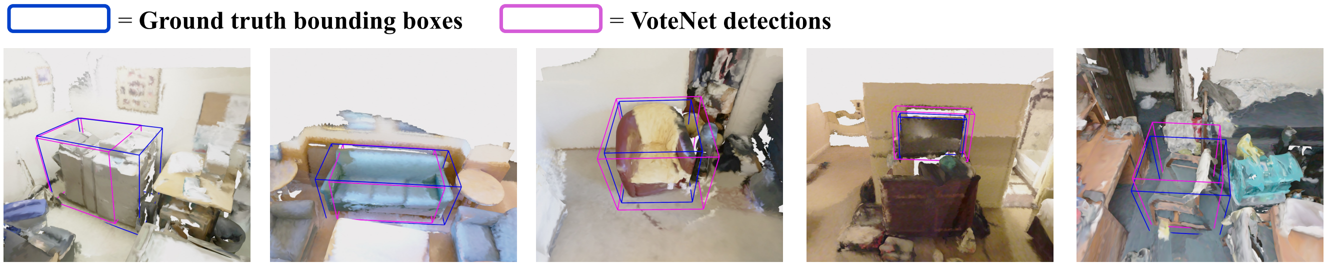


Figure 4. Comparison of ground truth bounding boxes (in blue) and VoteNet detections (in purple) used in the low guidance 3D visual grounding setting.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. ReferIt3D: Neural Listeners for Fine-grained 3D Object Identification in Real-world Scenes. In *ECCV*, pages 422–440. Springer, 2020. 1

[2] Eslam Bakr, Yasmeen Alsaedy, and Mohamed Elhoseiny. Look Around and Refer: 2D Synthetic Semantics Knowledge Distillation for 3D Visual Grounding. In *NeurIPS*, pages 37146–37158, 2022. 1

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. *NeurIPS*, 33:1877–1901, 2020. 1

[4] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language. In *ECCV*, pages 202–221. Springer, 2020. 1

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *CVPR*, pages 5828–5839, 2017. 1

[6] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. TransRefer3D: Entity-and-Relation Aware Transformer for Fine-Grained 3D Visual Grounding. In *ACM International Conference on Multimedia*, pages 2344–2352, 2021. 1

[7] Joy Hsu, Jiayuan Mao, and Jiajun Wu. NS3D: Neuro-Symbolic Grounding of 3D Objects and Relations. In *CVPR*, pages 2614–2623, 2023. 1

[8] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-View Transformer for 3D Visual Grounding. In *CVPR*, 2022. 1

[9] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom Up Top Down Detection Transformers for Language Grounding in Images and Point Clouds. In *ECCV*, pages 417–433. Springer, 2022. 1

[10] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep Hough Voting for 3D Object Detection in Point Clouds. In *CVPR*, pages 9277–9286, 2019. 1

[11] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. LanguageRefer: Spatial-Language Model for 3D Visual Grounding. In *CoRL*, pages 1046–1056. PMLR, 2022. 1