

# 1 **Supplementary Material for ROBOPoint:** 2 **A Vision-Language Model for Spatial Affordance Prediction for Robotics**

## 3 **A Overview**

4 This document provides additional details for model tuning (Sec. B), dataset generation (Sec. C) and  
5 results visualization (Sec. D) that support our main paper.

## 6 **B Instruction Tuning**

7 ROBOPoint is instruction-tuned from a Vicuna-v1.5-13B base model [1] with a ViT-L/14 336px  
8 image encoder pretrained with CLIP [2]. The projector is a 2-layer MLP pretrained on the 558K  
9 subset of the LAION-CC-SBU dataset with BLIP captions from [3]. The instruction tuning took 40  
10 hours on 16 A-100 GPUs with a batch size of 16 per-GPU. The learning rate is set to  $4e-5$ .

## 11 **C Data Generation**

12 Table A shows more examples from our procedurally generated synthetic dataset for object reference  
13 and free space reference.

14 We sample assets that one can find in a typical kitchen environments (e.g. dishwasher, hood, table,  
15 fridge) and use heuristics to place them in random, but semantic layouts in the scene. Once the  
16 furniture assets are added to the scene. We used a large object dataset sampled from ACRONYM  
17 [4]. Object positions are randomly sampled on support surfaces (e.g. countertop, table) and the  
18 orientations are determined by their stable poses. Poses that result in the object being in collision  
19 with the existing scene are rejected. We place cameras randomly in the scene and select those  
20 with at least three visible objects (visible means the number of points within segmentation mask  
21 is larger than 100) and at least 1 valid relationship between a pair of visible objects. The diverse  
22 view distribution allow ROBOPoint to maintain a consistent prediction across different viewpoints.  
23 Around 660K (image, relation) pairs are generated from 10K scenes.

24 We use the 3D bounding boxes of objects, surfaces and containers in the scene layout to compute a  
25 set of pairwise relations, including left, right, in front, behind, above, below, next to, on, inside, on  
26 left part, on right part, on front part, on back part. Note that although these relations are templated,  
27 the model fine-tuned on these data is able to generalize to new types of relations, as shown in  
28 Fig. A. For each relation, we first sample points on the object being referenced to create an example  
29 for object reference. Around 1 to 50 ground truth points are sampled per image. We convert the  
30 sampled points to a list of image coordinates normalized between 0 and 1 and use that as the ground  
31 truth response.

32 One caveat for these procedurally generated scenes is that the objects do not have rich text descrip-  
33 tions. Most objects just have a category name. We get around this problem by adding visual prompts  
34 to the rendered images. Specifically, we draw colored bounding boxes around the objects referenced  
35 in the language instruction. As a result, a typical instruction in the synthetic data will look like:  
36 “There is an object surrounded by a red rectangle in the image. Find some places in the free area to  
37 the left of the marked object.” Note that we do not add these visual prompts during testing, and thus  
38 do not require object detection. The idea is that the model learns to detect objects from other sources  
39 of data (e.g. LVIS [5]), and it will focus on relational reasoning when dealing with the object and  
40 space reference data.

## 41 **D Qualitative Examples**

42 Fig. A shows more qualitative comparisons of ROBOPoint against baselines on RoboRefIt [6] and  
43 WHERE2PLACE data, including examples demonstrating generalization to novel relation types and  
44 cases where ROBOPoint underperforms GPT-4o [7].

Relation	Above	Behind
		
Prompt	The image features an item encased in a red rectangular border. Locate several spots within the vacant space situated above the bordered item.	In the image, an object is framed by a red rectangle. Locate a few points on an object that is situated behind the framed object.
Relation	Between	Inside
		
Prompt	In the image, there is an item framed by a red rectangle and another item encased within a green rectangle. Locate several points upon the item situated between the two highlighted items.	The image depicts a container delineated by a red rectangular border. Pinpoint several spots within the vacant area enclosed by the outlined container.
Relation	Right	On left part
		
Prompt	The image features an object outlined by a red rectangle. Locate several points on an item that is situated on the right side of the marked item.	The image showcases an area demarcated by a red rectangle. Locate a few points within a vacant area on the right side of the marked surface.

Table A: Examples from the synthetic dataset used to teach ROBOPOINT relational object reference and free space reference. The red and ground boxes are visual prompts to indicate reference objects and the cyan dots are the visualized ground truth (not included in the image inputs to the model).

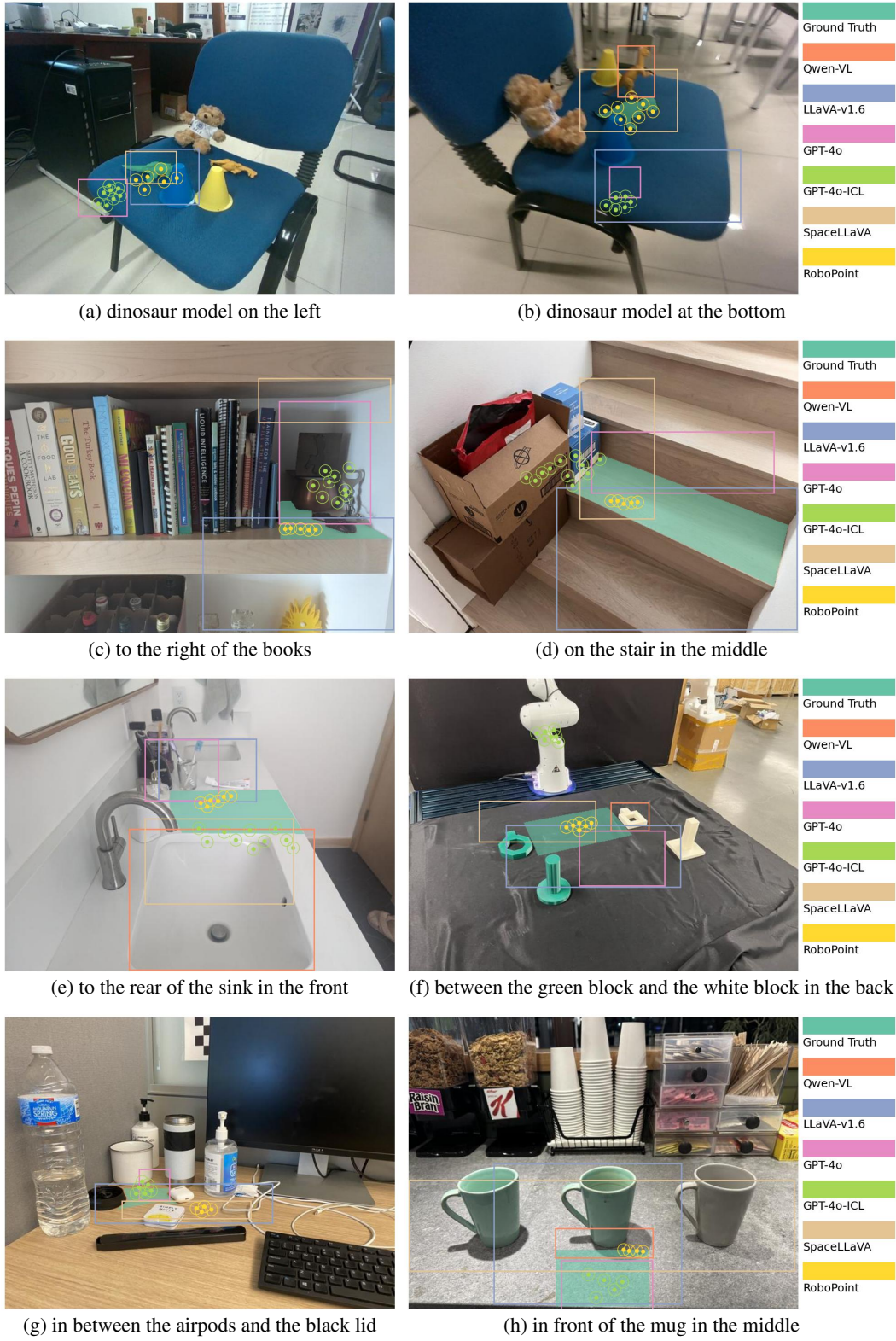


Figure A: Qualitative results on RoboRefIt (a, b) and WHERE2PLACE (c, d, e, f, g, h), including cases with relations unseen during training (d, e, f, h) and where GPT-4o performs better (g, h).

## References

- [1] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- [2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [3] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023.
- [4] C. Eppner, A. Mousavian, and D. Fox. Acronym: A large-scale grasp dataset based on simulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6222–6227. IEEE, 2021.
- [5] A. Gupta, P. Dollar, and R. Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.
- [6] Y. Lu, Y. Fan, B. Deng, F. Liu, Y. Li, and S. Wang. Vl-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 976–983. IEEE, 2023.
- [7] OpenAI. Hello gpt-4o, May 2024. URL <https://openai.com/index/hello-gpt-4o>.