Supplementary Material of LanguageRefer: A Spatial-Language Model for 3D Visual Grounding

Junha Roh, Karthik Desingh, Ali Farhadi, Dieter Fox
Paul G. Allen School, University of Washington, United States
{rohjunha, kdesingh, ali, fox}@cs.washington.edu

In this document, we provide additional details and results of the proposed method. This document is accompanied by a folder containing our code implementation and a video explaining our method. We also created a webpage to provide evaluation results: https://sites.google.com/view/language-refer.

1 Examples of ReferIt3D [1] Dataset

ReferIt3D [1] proposed a 3D visual grounding task with the dataset with referring language annotation on a 3D reconstructed indoor scene dataset, ScanNet [2]. Following the official splits of ScanNet [2], ReferIt3D [1] provided natural language annotation of referring one of 76 target classes which is called Nr3D. It also provides template-based referring language on spatial relationship, Sr3D. Based on the datasets, it proposed the 3D visual grounding task. From a scene, ground-truth bounding-boxes, corresponding pointclouds, and the language description of the target object are given. The goal of the task is to choose one bounding-box from candidates.

We visualized three example scenes (scene0011_00, scene0231_00, scene0141_00) with bounding-boxes, utterances, and corresponding target bounding-boxes in Figure 1-3. Subfigures a show all bounding-boxes (red) in the example scenes. Bounding-boxes of selected target classes are highlighted in blue. For instance, in Figure 1 (b-c) two bounding-boxes of tables are in blue since we choose table as the target class for visualization. Rest of figures (Figure 1 (b-c), 2 (b-e), 3 (b-d)) show the individual bounding-box of objects (yellow) in the target classes. We add virtual robot paths for the application of robot navigation. Red dots indicate the random starting positions and yellow dots demonstrate paths to reach the target object. Each caption contains an example utterance for the corresponding target object.

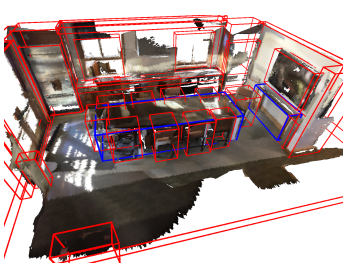
2 LanguageRefer at the Inference Stage

Figure 4 shows the inference stage of the proposed method. At inference, we followed the approach of InstanceRefer [3] to filter out objects that do not belong to the predicted target class. A target classifier takes the language utterance as input and predicts the target class. Filtering masks are generated by comparing the predicted target class to predicted class labels from the semantic classifier. In order to reduce the chance of removing the true target instance in the filtering process, top-k class predictions (from the semantic classifier) for each object are compared to the predicted target class (not shown in the figure). Filtering masks are applied to the output embeddings of the spatial-language model to refine objects only related to the predicted target class.

3 Qualitative Results of LanguageRefer

Figure 5 shows the qualitative prediction result of LanguageRefer on the first example scene (scene0011_00) with natural language utterances. In Figure 5 (a) and (b), the model correctly chooses the target object given utterances in the test dataset as well as the custom utterances such as "***a smaller table***." Figure 5 (c) shows the failure case where the custom utterance "***table without any chairs around***." is given but the model selected another table at the center of the room. Expressions

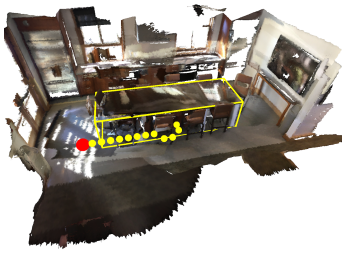
5th Conference on Robot Learning (CoRL 2021), London, UK.



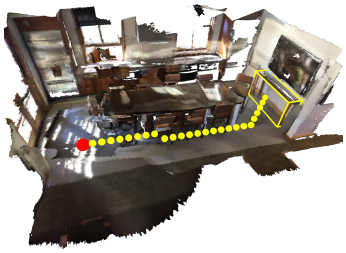
(a) Two bounding-boxes of tables are shown in blue while the bounding-boxes of other classes in red.

(b) A Yellow bounding-box shows the ground-truth bounding-box with the utterance of "***this is the large conference table with many chairs.***."

(c) A Yellow bounding-box shows the ground-truth bounding-box with the utterance of "***the desk directly below the board on the wall***."



(d) A virtual robot path from a random position in the scene to the target object from (b) is shown in yellow dots.



(e) A virtual robot path from a random position in the scene to the target object from (c) is shown in yellow dots.

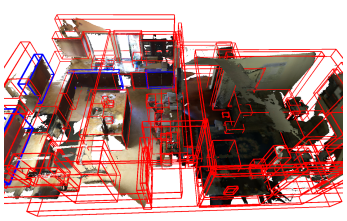
Figure 1: **Examples of ReferIt3D [1] dataset.** Figure (a) shows all the bounding-boxes (red) in an example scene with two highlighted bounding-boxes (blue) of the target class. Figure (b-c) show two utterance examples and corresponding target object bounding-boxes. Figure (d-e) visualize virtual robot paths generated by A* from a random position to two target objects. Yellow dots indicate the path and red dots indicate the (random) initial positions.

such as without seem to be rare; our model correctly predicts all referred tables from corresponding utterances in the dataset:

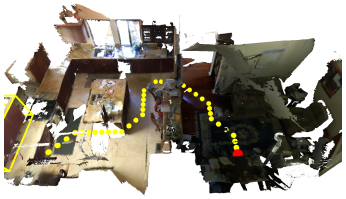
"***this is the large conference table with many chairs***", "***the desk directly below the board on the wall***", "***the biggest table in room***", "***the large table in the middle of the room***", "***the thin wooden table underneath the television and immediately to the left of the trash can***", "***smaller table against the wall***", "***select the long table in the middle of the room***", "***choose the table that is up against the wall***", "***the long conference table in the middle of the room***", "***choose the table that sits against the wall***", "***the largest table in the room***", "***select the table underneath the tv***", "***a small table below the television on the wall***", "***the very big dining table in the center of the room***."

In Figure 6, we asked the model to choose one of the three stacked boxes with natural language utterances. The model failed to select the top box (in yellow) and selected the box in the middle (in red) in Figure 6 (a). When we replace predicted class labels from PointNet++ [4] by the ground-truth class labels, the model was able to choose the correct box on top in 6 (b). However, the attempts to select the box in the middle failed with or without ground-truth class labels in Figure 6 (c). Figure 6 (d) shows the successful reference of the bottom box by the model with predicted class labels. Now the robot paths are visualized with the color of prediction; green if correct, red otherwise.

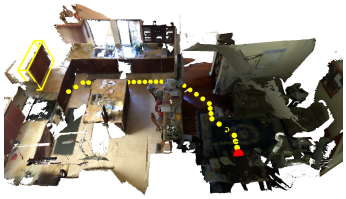
We examined the predicted class labels of three boxes: microwave, box, box (from top to bottom). This caused an incorrect reference of the top box: 0/7. After fixing the incorrect class label by the ground-truth class label, the accuracy of the reference task of the top box got higher: 6/7. However, the ground-truth class label did not improve the accuracy of the reference task of the box in the middle: 0/6. The reference task of the bottom box was 6/6 before fixing the label. By providing ground-truth labels, we were able to disentangle reference errors from perception errors. From the three-box example, we found the model was not able to refer to the box in the middle of the vertical stack.



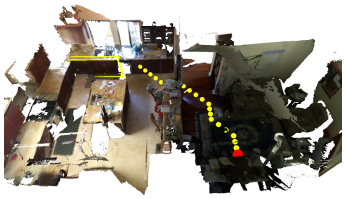
(a) Four bounding-boxes of kitchen cabinets are shown in blue and the other bounding-boxes are shown in red.



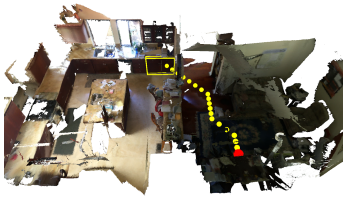
(b) A Yellow bounding-box shows one of the bounding-box of kitchen cabinets with the utterance of "***Kitchen cabinet to the left of the stove***."



(c) A Yellow bounding-box shows one of the bounding-box of kitchen cabinets with the utterance of "***This upper cabinet is between the stove and sinks***."



(d) A Yellow bounding-box shows one of the bounding-box of kitchen cabinets with the utterance of "***The cabinets under the sink***."



(e) A Yellow bounding-box shows one of the bounding-box of kitchen cabinets with the utterance of "***Of the two bottom cabinets, choose the one on the right***."

Figure 2: **Examples of ReferIt3D [1] dataset.** Figure (a) shows all the bounding-boxes (red) in an example scene with four highlighted bounding-boxes (blue) of kitchen cabinets. Figure (b-e) show four utterance examples, corresponding target object bounding-boxes, and robot paths to reach the objects. Yellow dots indicate the path and red dots indicate the (random) initial positions.

In Figure 7, we evaluated the accuracy of the reference task of kitchen cabinets with natural language utterances. Figure 7 (a) and (b) show successful examples of reference. Figure 7 (c) shows a failure case.

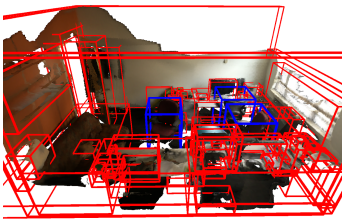
Note that the top-5 class predictions of the kitchen cabinets are noisy:
top-box: ['cabinet', 'cabinets', 'kitchen cabinets', 'bathroom cabinet', 'kitchen cabinet'],
middle-box: ['cabinet', 'bathroom cabinet', 'cabinets', 'kitchen cabinet', 'kitchen cabinets'],
bottom-box: ['kitchen cabinets', 'cabinet', 'cabinets', 'kitchen cabinet', 'bathroom cabinet'].

It shows that our model is robust with subtle changes in class labels. Even without a unified format of similar classes including plurals, our model was able to accurately refer to the correct objects. We do not preprocess utterances except for tokenization; preprocessing of language expressions or transforming the utterance into a fixed form is not used [5].

4 Qualitative Comparison to ReferIt3D [1]

We have compared the proposed method to ReferIt3D [1] and Figure 8 (a-g) show examples of predictions (on scene0699_00) with corresponding utterances in captions. Figure 8 (a) shows some highlighted objects in the scene. The scene has a bed at the bottom of the image, a walk-in closet on top, a desk on its right. Blue bounding-boxes highlight bags and yellow bounding-boxes show backpacks. We examine utterances that select one of the bags in the bedroom.

Figure 8 (b) and (c) show results from LanguageRefer and ReferIt3D [1] respectively with the utterance "***The light brown bag on the floor closest to the bed***." The proposed method correctly selected the ground-truth bag (in green, in Figure 8 (b)) while ReferIt3D [1] chose an incorrect bag (in red, in Figure 8 (c)).



(a) Three bounding-boxes of chairs are shown in blue and the other bounding-boxes are shown in red.

(b) A Yellow bounding-box shows one of the bounding-box of chairs with the utterance of "***Chair closest to the door***."

(c) A Yellow bounding-box shows one of the bounding-box of chairs with the utterance of "***select the chair pushed in at the desk***."

(d) A Yellow bounding-box shows one of the bounding-box of chairs with the utterance of "***The chair is the red one closest to the window facing towards the blue ball and away from the desk***."

Figure 3: **Examples of ReferIt3D [1] dataset.** Figure (a) shows all the bounding-boxes (red) in an example scene with three highlighted bounding-boxes (blue) of chairs. Figure (b-d) show three utterance examples, corresponding target object bounding-boxes, and robot paths to reach the objects. Yellow dots indicate the path and red dots indicate the (random) initial positions.

Figure 8 (d) and (e) show results with the utterance "***It is the bag against the wall, not in the closet***." While our approach successfully chose the intended bag, ReferIt3D [1] chose the backpack nearby instead of choosing a bag. Confusion with objects of other classes often happened from ReferIt3D [1].

Figure 8 (f) and (g) show results with the utterance "***the bag in the closet next to the cardboard box***." For the utterance of choosing a bag in the closet, both methods failed. They chose the same objects as with the utterance "***It is the bag against the wall, not in the closet***." We found that for all the utterances to select the bag in the closet, both methods never chose the intended bag. It happened when we provide the ground-truth class labels to our method and it was not caused by perception failure.

5 Orientation Annotation for View-Dependent Utterances

In addition to the proposed model, we have collected orientations of the view-dependent utterances. View-dependent utterances without information about the original viewpoint make the reference task challenging. For instance, utterances such as "The door is wood with the handle on the left side." assume specific orientations of the agent and it is impossible to recover the true orientation without knowing the referred object, not like view-dependent utterances with explicit view-point information such as "Facing the foot of the bed." However, the original dataset of ReferIt3D [1] does not distinguish the utterances without orientation information from those with orientation information. Therefore, we split the view-dependent (VD) utterance category into two subcategories, ***VD-explicit*** and ***VD-implicit***, where VD-explicit has explicit view-point information in the utterance. Then we collected orientations that make the utterances valid from human annotators.

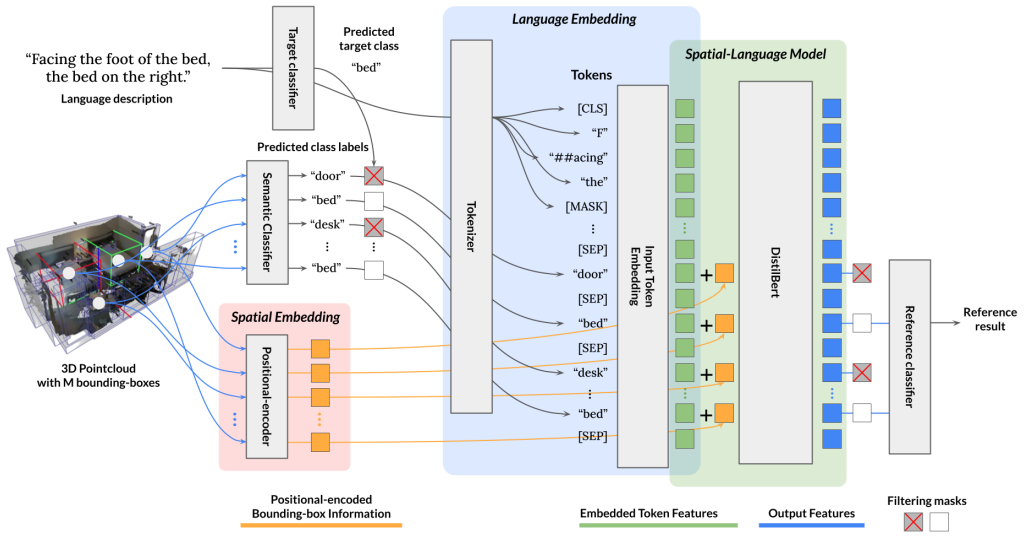
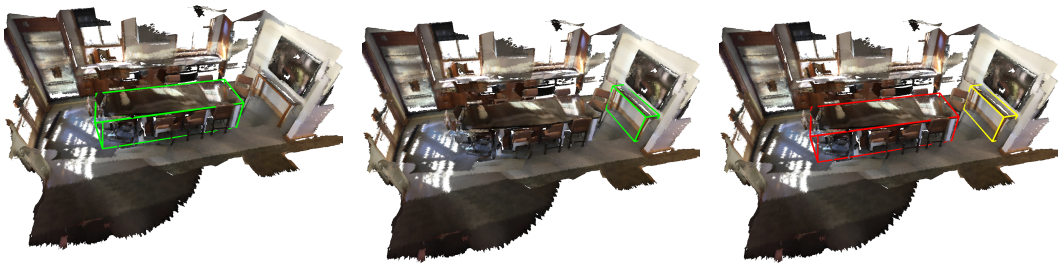


Figure 4: **Detailed overview of LanguageRefer at the inference stage.** At inference, an extra target classifier is employed to exclude objects that are not related to the target class in the reference task. Please refer to the details in the section 2.



(a) LanguageRefer successfully selects the referred table at the center with a custom utterance "***table with lots of chairs***." as well as the utterance from the dataset "***this is the large conference table with many chairs***."

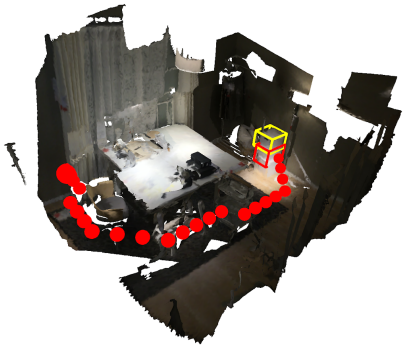
(b) LanguageRefer chooses the correct table on the side with a custom utterance "***a smaller table***." as well as the utterance from the dataset "***the desk directly below the board on the wall***."

(c) LanguageRefer chooses the incorrect table at the center (in red) when the desired target object is the smaller table (in yellow) with a custom utterance "***table without any chairs around***."

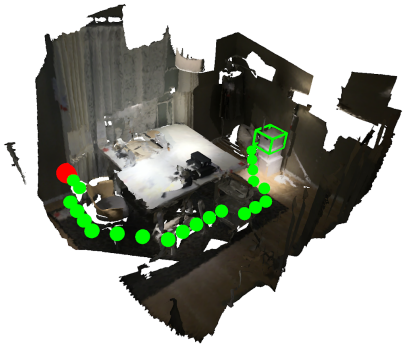
Figure 5: **Qualitative analysis of LanguageRefer result with table as the target class.** Figure 5 (a) and (b) show successful results of LanguageRefer references in an example scene (green) and input utterances. The proposed model predicts correct target objects with custom utterances. Figure 5 (c) shows an failure case of the proposed method with a custom utterance.

We set four standard orientations assuming the agent is in the room (around the center of the scene) and ask the annotators to select all of the orientations that can be considered valid from the utterance. Figure 10 shows examples of four orientations. With the assumption of the agent being inside of the room, we found that four orientations are good enough to recover the original viewpoints of the speakers. In total, 12,680 view-dependent utterances of the Nr3D dataset were annotated. From those, 5,942 utterances are classified as VD-explicit. For train and test split, 10,206 and 2,474 utterances were annotated.

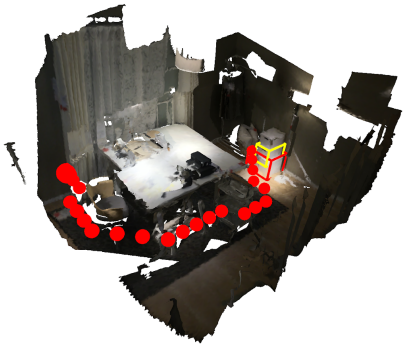
We also provide a link to the orientation annotation webpage (Orientation Annotation Webpage). It is recommended to use a computer or a laptop to view the page. Please do not click the view-annotation checkboxes (it is still alive so clicks can affect the actual annotation data.) The first page shows the links to all the scenes for annotation. If you click one scene, you can see the list of utterances with some flags. By clicking a single utterance, all the bounding-boxes with highlighted target bounding-boxes are rendered. A green bounding-box is the ground-truth target and the red bounding-boxes are distractors that belong to the target class but not the target object in-



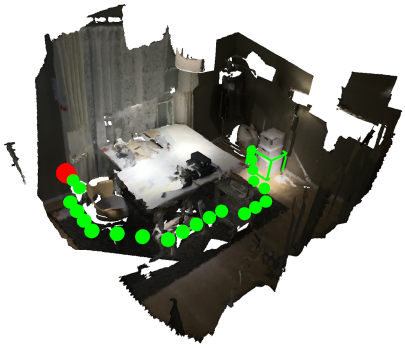
(a) LanguageRefer failed to select the top box from the utterance "***of the three boxes stacked pick the top one***." A yellow bounding-box shows the ground-truth target object and a red bounding-box shows the incorrectly selected target object.



(b) When we provide the ground-truth class labels to the model, it chooses the correct box from the same utterance as in Figure 6 (a).



(c) LanguageRefer chooses the incorrect box at the bottom (in red) with the utterance "***middle box in the stack***." indicating the ground-truth box in the middle (in yellow).

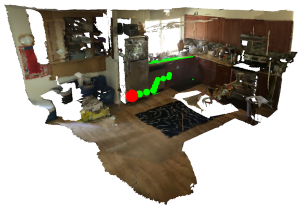


(d) LanguageRefer chooses the correct box at the bottom (in green) with utterances "***the bottom box in the stack of three boxes***." and "***the large white box on the bottom***."

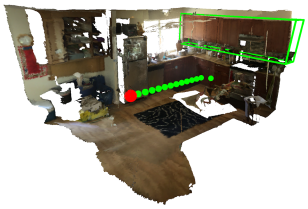
Figure 6: **Qualitative analysis of LanguageRefer result and effect of predicted class labels.** Figure 6 (a) shows a failure case of selecting the top box from the stack of boxes and Figure 6 (b) shows the corrected reference when the ground-truth class labels are provided. Figure 6 (c) and (d) shows a failure and a successful case of different boxes, respectively. Ground-truth boxes are shown in yellow, correct and incorrect guesses are shown in green and red, respectively.

stance. Then press any number in {1, 2, 3, 4} on the keyboard. You can see the scene with the canonical orientation of your selection. You can zoom in/out, translate, rotate with your mouse. We collect annotations on utterances only with correct guesses from humans and mention the target class. Figure 9 shows examples of four standard orientations and Figure 10 shows the annotation interface.

Note that, in the process of ReferIt3D [1] annotation, the ground-truth target class and the distinguished bounding-boxes of the target class are provided to the annotators. We also provide those information to the annotators. However, in the actual task of ReferIt3D [1], the model does not have access to the target class and many other bounding-boxes from other classes are given as you can see in Figure 1 (a). If some utterances are assuming the shared view or orientation of speakers, the ambiguity can be easily resolved by human listeners since they have extra information and they can manipulate orientation as well. However, the same assumption can make the reference task even more challenging because the model needs to verify some hypothetical orientations with uncertainty of classes among multiple candidates.



(a) LanguageRefer successfully selects the referred kitchen cabinets with the utterance "***a double sink is on top of this section of low cabinets***."



(b) LanguageRefer chooses the correct kitchen cabinets with the utterance "***higher section of kitchen cabinets***."

(c) LanguageRefer chooses the incorrect kitchen cabinet on the left side (in red) when the desired target cabinets are in the middle (in yellow) with the utterance "***lower cabinets next to stove***."

Figure 7: **Qualitative analysis of LanguageRefer result with kitchen cabinets as the target class.** Figure 7 (a) and (b) show successful results of LanguageRefer references in an example scene (green) and input utterances. The proposed model predicts correct target objects with custom utterances. Figure 7 (c) shows an failure case of the proposed method.

6 Ablation of Positional Encoding and DistilBert

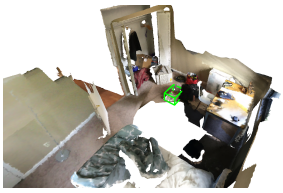
In order to examine the effect of the spatial encoding (sinusoidal positional encoding function) [6] and the base language model (DistilBert [7]), we have trained ablations models.

In the first ablation model, we replaced the sinusoidal positional encoding function with a linear layer to transform a 6-dimensional input vector of bounding-box information to a 768-dimensional embedding vector. In the second ablation model, we keep the spatial encoding to the positional encoding function but replaced DistilBert [7] with BERT [8].

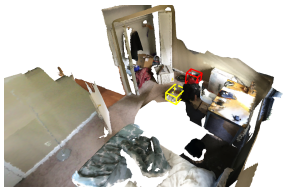
The first ablation model of the positional encoding achieved 37.8 % accuracy on Nr3D while our final model achieved 43.9 % on Nr3D. It shows that selection of an effective spatial encoding scheme such as the sinusoidal is important.

The second ablation model with BERT as the base language model achieved 45.3 % accuracy on Nr3D which is slightly higher than the accuracy of the model with DistilBert [7] (43.9 %). However, when it was trained on Sr3D, it only achieved 49.3 % while the original model achieved 56.0 %. We observed instability in training with the BERT [8] model. We also observed that the successfully trained model on Nr3D converged faster than the DistilBert-based model. During the training, the total loss of the model on Sr3D surged in the middle and did not recover. While these are some observations during our ablation, the second study with BERT [8] model is inconclusive. We chose DistilBert-based model in our framework is because it is lightweight and easy to train. Our goal is to develop a modular approach that can be easily modified based on the advancements in the area of learning embeddings, especially in NLP and computer vision.

(a) Target objects (bag) in blue and non-target objects (backpack) in yellow.

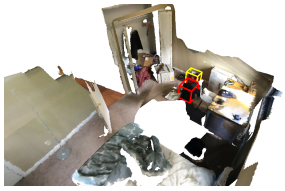


(b) LanguageRefer chooses the correct bag (in green) with the utterance "***The light brown bag on the floor closest to the bed***."

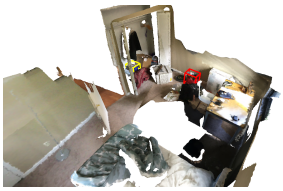


(c) ReferIt3D [1] chooses an incorrect bag (in red) with the utterance "***The light brown bag on the floor closest to the bed***." The ground-truth bag is in yellow.

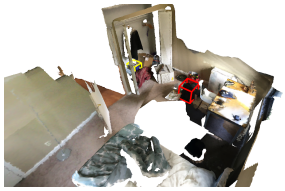
(d) LanguageRefer chooses the correct bag (in green) with the utterance "***It is the bag against the wall, not in the closet***."



(e) ReferIt3D [1] chooses an incorrect object, backpack (in red), instead of the ground-truth bag (in yellow) with the utterance "***It is the bag against the wall, not in the closet***."

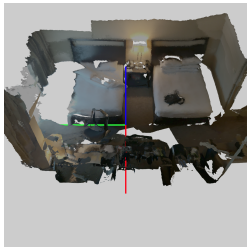


(f) LanguageRefer fails to choose the correct bag (in yellow) but chooses an incorrect bag (in red) with the utterance "***the bag in the closet next to the cardboard box***."

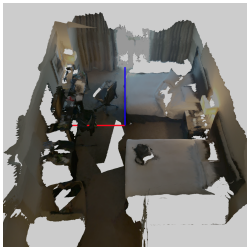


(g) ReferIt3D [1] chooses an incorrect backpack (in red) again, instead of the ground-truth bag (in yellow) with the utterance "***the bag in the closet next to the cardboard box***."

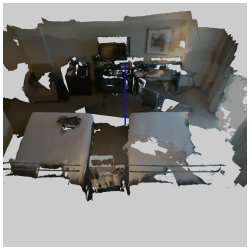
Figure 8: **Results of LanguageRefer and ReferIt3D [1] on an example scene.** Figure 8 (a) shows an example scene with highlighted object bounding-boxes of two classes: target objects (bag) in blue and non-target objects(backpack) in yellow. Figure 8 (b-g) show three utterance examples and corresponding reference results of LanguageRefer and ReferIt3D [1].



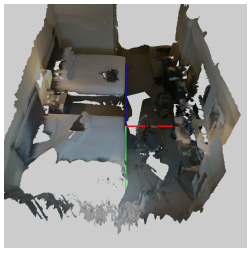
(a) An example of the standard orientation 1.



(b) An example of the standard orientation 2.



(c) An example of the standard orientation 3.



(d) An example of the standard orientation 4.

Figure 9: **Examples of standard orientations for view-point annotation (a-d).** We assume that the robot is always inside of the room except for the cases specified by utterances.

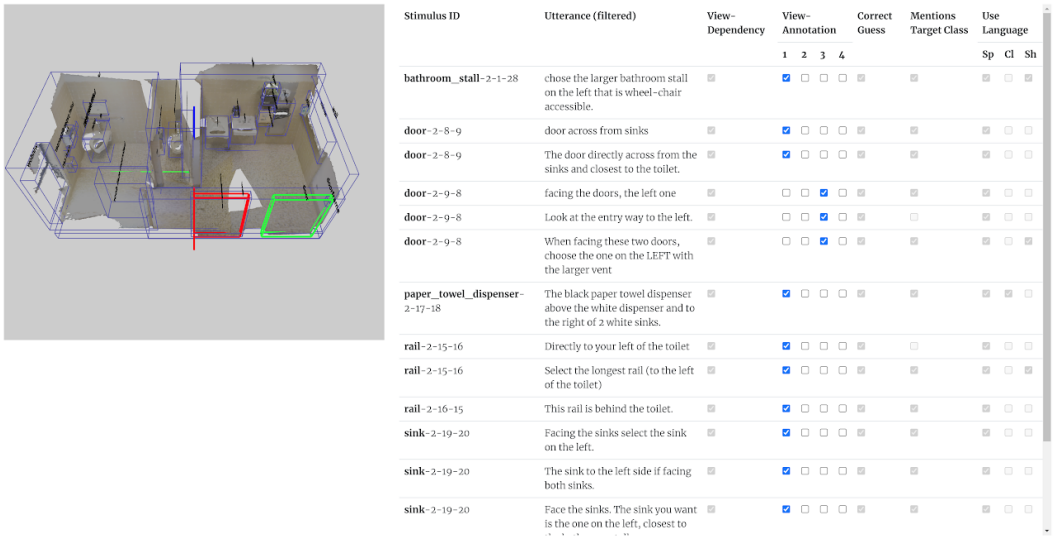


Figure 10: **Interface of the orientation annotation.** On the left side, a 3D visualization of the scene with overlaid bounding-boxes and class labels is provided. On the right side, a table of utterance information with orientation annotation checkboxes is shown. By clicking each row in the table, the highlights of the bounding-boxes in 3D visualization is changing with respect to the clicked utterance. A green bounding-box is the true target object while red bounding-box(es) is the distractor object in the same class. Flags 'Correct Guess' and 'Mentions Target Class' indicate whether utterances are considered valid (according to the official ReferIt3D [1] evaluation). Flags 'View-Dependency' and 'Use Language' provide information about the utterances. 'Sp', 'Cl', 'Sh' indicate whether the utterance is using spatial relationship, color, and shape, respectively. For instance, the utterance "The black paper towel dispenser above the white dispenser and to the right of 2 white sinks." uses color ("black" and "white") and spatial relationship ("above", "to the right of").

References

[1] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. *16th European Conference on Computer Vision (ECCV)*, 2020.

[2] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

[3] Z. Yuan, X. Yan, Y. Liao, R. Zhang, Z. Li, and S. Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring, 2021.

[4] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 2017.

[5] M. Feng, Z. Li, Q. Li, L. Zhang, X. Zhang, G. Zhu, H. Zhang, Y. Wang, and A. Mian. Free-form description guided 3d visual graph network for object grounding in point cloud, 2021.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[7] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.