Supplementary of Viewpoint-Aware Visual Grounding in 3D Scenes

Epression Generation Given a relation tuple, e.g ("chair", "left", "bed"), we extend it to a sentence with the following template:

⟨Target Phrase⟩ ⟨Relation Phrase⟩ ⟨Mentioned Phrase⟩

The ⟨Target Phrase⟩ ("chair") and ⟨Mentioned Phrase⟩ ("bed") represent the phrases associated with the target and mentioned objects. These phrases begin with "the", "a", "this", or "that" followed by the object classes. ⟨Relation Phrases⟩ are sampled from a predefined pool based on the predicted relation ("left"). Then, the generated sentences are concatenated into an expression. When the expression consists of multiple short sentences, we introduce a random probability for replacing the ⟨Target Phrase⟩ with "it" starting from the second sentence onward. This enhancing diversity in the construction of the expressions. More example of synthetic expressions are shown following:

- This chair is left to a picture.
- The chair is left to another chair. It is in front of a cabinet. It is behind another cabinet.
- The cup is to the right of a picture. It is left to a door. It is behind the backpack. It is above the desk.
- This ladder is to the left of an umbrella. It is behind a shelf.
- A lamp is left to a pillow. It is behind the curtain.
- A towel is left to another towel. It is in front of the cabinet.
- The desk is to the right of the chair. It is in front of the couch. It is behind another desk.
- The fan is to the left of the chair. It is in front of another chair. It is behind a door.
- That chair is on the left side of another chair. It is in front of another chair. It is behind another chair.
- A backpack is right to the door. It is in front of the door. It is behind a refrigerator.
- That sink is to the left of the toilet. It is behind the toilet.
- This easel is to the left of the chair. It is behind the chair.
- This table is on the right side of the stove. This table is in front of the stove. It is behind the chair.
- A toilet is behind the light.
- That printer is on the left side of another printer. The printer is in front of the cabinet. It is behind the door. It is above the table.
- That monitor is left to the picture.
- The picture is in front of the dresser.
- A window is on the left side of the whiteboard.
- This couch is to the left of this chair. It is in front of the chair. It is behind another chair.
- This monitor is on the left side of this pillow. It is in front of the couch.

Effectiveness of Viewpoint Prediction. To study the effectiveness of the viewpoint predictor, we conducted an analysis of pretrained model variations within the viewpoint prediction task. The results are presented in Table 1, where (1) serves as a baseline representing random guessing, (5) corresponds to the model trained with Curriculum Filtering and all synthetic training data, and (6) is the model trained with viewpoint data augmentation based on (5).

When comparing models trained with different proportions of synthetic data ((2-3) and (5)), it is notable that enlarging the training set size correlates positively with improved performance in viewpoint prediction. Additionally, incorporating the Curriculum Filtering mechanism (4-5) yields further enhancements, manifesting as a 3% and 1% increase in recall for perspective and location, separately. Meanwhile, it benefits visual grounding task, as shown in Table 4 of the main paper. We also experimented with longer (1.5x) curriculum durations but these resulted in worse performance (39.2% of Acc@0.50).

Contribution of Viewpoint Data Augmentation. Viewpoint data augmentation provides more data based on different views in the same scene to enhance the viewpoint prediction and object representation. Shown in Row 6 of Table 1, with the viewpoint data augmentation the model boosts perspective prediction accuracy from 40% to 43%, but does not affect the location prediction (84% for both).

Contribution of Uniform Object Representation. UOR contributes to the model by providing more robust object features. To support this, we conduct an experiment on our VPPNet with different loss weights of α_2 and α_3. In our experiments, we set $\alpha_2 = \alpha_3$. From Table. 2, we can find that the higher parameter benefits the model before $\alpha = 16$,

	Model(%)					Metrics(%)	
	Data Aug.	Curr. Filt.	25%	50%	100%	Perspective	Location
1						13%	69%
2		✓	✓			31%	78%
3		✓		✓		37%	82%
4					✓	37%	83%
5		✓			✓	40%	84%
6	✓	✓			✓	43%	84%

Table 1. Recall of the viewpoint prediction in synthetic validation set. 25%, 50% and 100% are the proportions of the synthetic data we used to train the model.

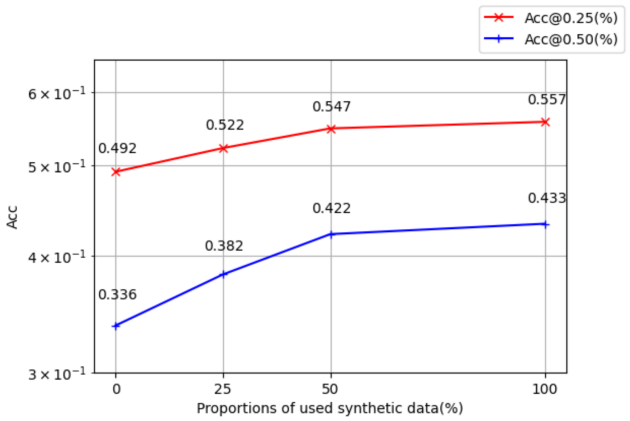

Figure 1. VPP-Net trained with different proportions of synthetic data. Horizontal axis represents the proportions of the synthetic data we used during the training. Vertical axis represents the grounding accuracy.

α_2	Acc@0.25 %	Acc@0.50 %
0	53.8	39.0
4	53.2	41.0
8	**55.6**	**43.3**
16	54.5	40.3

Table 2. Visual Grounding Result with different α_2 in Scanrefer. $\alpha_2 = 8$ is the hyper-parameter we used in our best model.

suggesting suggesting UOR regularization improves representation learning for grounding until it overpowers other objectives.

Proportions of Synthetic Data for Training We investigate the impact of varying proportions of the synthetic dataset within the ScanRefer [1] dataset. In our experimental setting, all modules in VPP-Net are retained. The models are trained individually with 25%, 50%, and 100% of synthetic data and evaluated on the ScanRefer Validation set. For the 0% data case, we follow the variant (5) in the ablation study. The performance is reported in terms of Acc@0.25 and Acc@0.50. As depicted in Fig. 1, we can see that an increase in the amount of synthetic data leads to improved results, enhancing performance in both Acc@0.25 and Acc@0.50 metrics.

Visualization of successful results in ScanRefer Besides the 4 examples we show in the main paper, we visualize more successful examples in Fig 2.

References

[1] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020. 2, 3

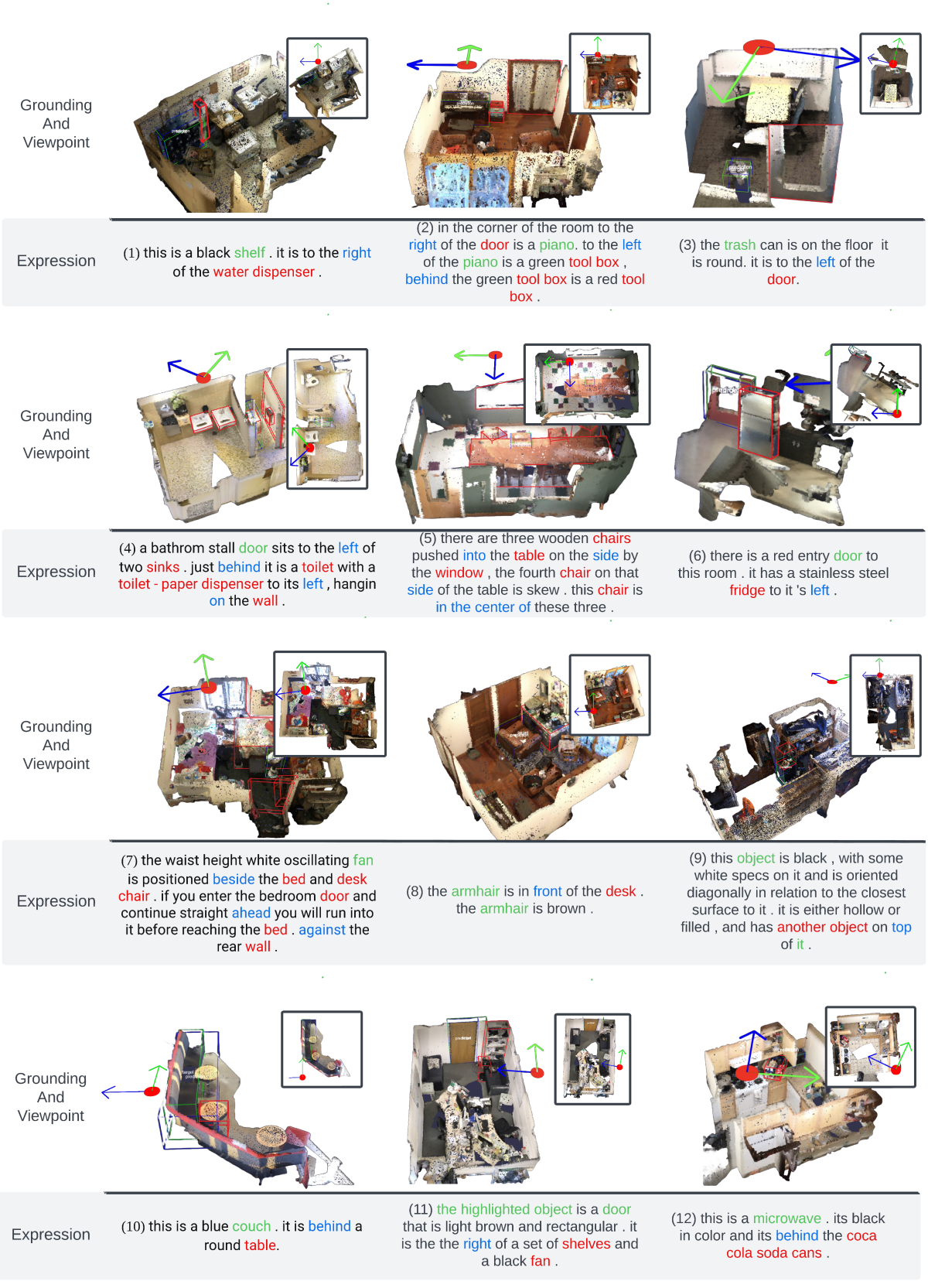

Figure 2. Exemplary examples of VPP-Net on Scanrefer [1]. The predicted observer position (red dot), facing direction (facing away from the green arrow), and 'right" (blue arrow) direction are shown in 3D and a top-down view (top right corner). The ground truth bounding boxes and target words are noted with green and the mentioned objects are noted with red. We also provide the predicted object bounding box in the image, shown in blue. The spatial relations are noted with blue in the text.