Question Generation to Disambiguate Referring Expressions in 3D Environment

Fumiya Matsuzawa, Ryo Nakamura, Kodai Nakashima, Yue Qiu, Hirokatsu Kataoka, Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST)

1. Overview

Our paper presents a novel task and method for question generation, aimed to disambiguate referring expressions within 3D indoor environments (3D-REQ). Referring to objects using natural language is a fundamental aspect of human communication, and an essential capability for robots in various applications such as room organization. However, human instructions can sometimes be ambiguous, which poses challenges to existing research on visual grounding in 3D environments that assumes referring expressions can uniquely identify objects.

To address this issue, we introduce a method inspired by human communication, where ambiguities are resolved by asking questions. Our approach predicts the positions of candidate objects that satisfy given referring expressions in a 3D environment and generates appropriate questions to narrow down the target objects. To facilitate this, we have constructed a new dataset (3D-REQ), containing input referring expressions with ambiguities and point clouds and output bounding boxes of candidate objects and questions to eliminate ambiguities. To our knowledge, 3D-REQ is the first effort to tackle the challenge of ambiguous referring expressions in 3D object grounding.

2. Proposal dataset (3D-REQ)

In the task of visual grounding by referring expressions in a 3D environment, datasets can be categorized into two types: those collected from human utterances, such as ScanRefer [1] and Nr3D [2], and those generated from 3D information using a rule-based method, like Sr3D [2]. Existing datasets typically contain only referring expressions that uniquely identify objects, without ambiguity. However, real-world referring expressions often include ambiguity, so it's crucial to account for this when developing practical applications.

To address the issue of ambiguity in referring expressions, we propose constructing a dataset that includes appropriate natural language questions, which help eliminate the ambiguity in each 3D scene. This dataset contains referring expressions, bounding boxes for all candidate objects that satisfy the referring expression, and questions asking about the feature or relationship causing the ambiguity, such as "What color is it". By incorporating these questions, the dataset aims to capture and handle the ambiguous representations that are common in real-world referring expressions.

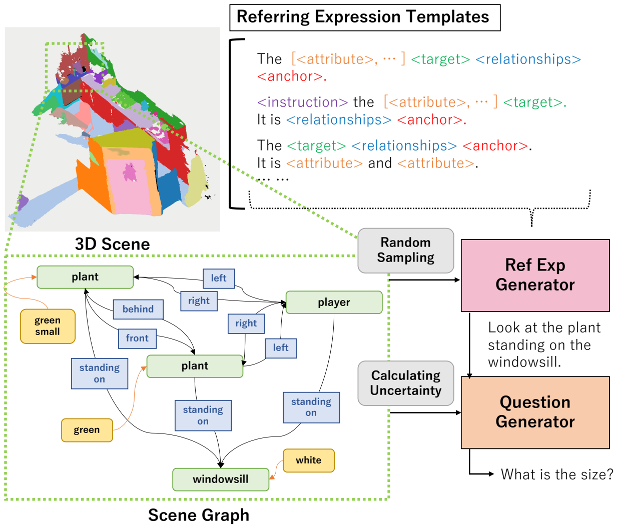

Referring Expression Templates
The [<attribute>, ···] <target> <relationships> <anchor>.
<instruction> the [<attribute>, ···] <target>. It is <relationships> <anchor>.
The <target> <relationships> <anchor>. It is <attribute> and <attribute>.
··· ···
3D Scene
Random Sampling
Ref Exp Generator
plant
left
right
player
behind
green small
right
front
left
standing on
plant
standing on
standing on
green
white
windowsill
Calculating Uncertainty
Look at the plant standing on the windowsill.
Question Generator
What is the size?
Scene Graph

Figure 1. Dataset generation process of the proposed dataset 3D-REQ.

The 3DSSG dataset [3], which serves as the foundation for our proposed dataset, is a large-scale indoor 3D dataset consisting of 3D meshes and densely annotated with 3D scene graphs. It provides detailed information about object attributes and relationships within the 3D environment, making it an ideal resource for generating referring expressions and addressing ambiguity.

During the dataset generation process, as shown in Figure 1, we create pairs of referring expressions and questions by randomly sampling attributes and relationships from the 3DSSG dataset. This allows us to consider the ambiguity of object identification within the 3D environment. We calculate the expected value of ambiguity for each potential question, aiming to select a question statement that would minimize indeterminacy after the question is asked. If a selected question successfully eliminates the ambiguity, the pair of question text and referring expression is added to the

Table 1. Evaluation of different model ablations on question generation and object detection applied to the proposed dataset.

Input	Query	Question generation				Ref Expression	
		BLEU	CIDER	METEOR	ROUGE	mAP	mAR
Geo	32	38.2	228.1	25.9	67.5	11.2	34.4
Geo	64	36.4	197.2	25.7	65.7	**11.8**	42.6
Geo	128	28.4	163.3	21.5	43.9	7.4	43.7
Geo+RGB	32	35.5	218.5	24.4	65.7	9.9	35.4
Geo+RGB	64	**41.9**	**239.9**	**26.9**	**69.4**	10.6	**49.4**
Geo+RGB	128	26.0	130.5	20.3	49.9	5.6	38.7

dataset. This approach helps create a more robust dataset that addresses the ambiguities often found in referring expressions, making it more applicable to real-world scenarios.

3. Experiments

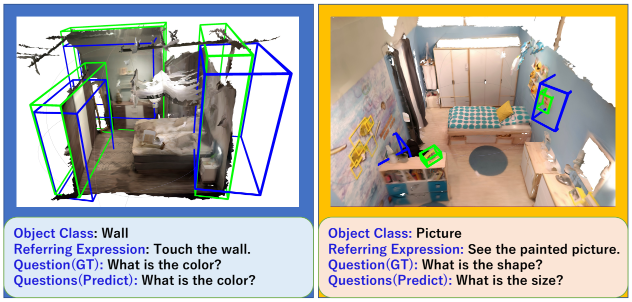
Experimental Setup: In summary, we evaluated our model using the 3D-REQ dataset. Our goal was to find bounding boxes for all candidate objects satisfying the referring expressions while generating questions to eliminate ambiguities. We implemented a question and localization network based on 3DETR [4], which is an object detector, and set the layer and head numbers to two for all transformer components. The model was trained with specific loss weights and a learning rate of 0.0001 for 40 epochs. We conducted ablation studies on the input type and query number, comparing 3D and 6D point clouds and testing with 32, 64, and 128 queries. For evaluation, we used mAR and mAP metrics for bounding box accuracy and BLEU [5], CIDER [6], METEOR [7], and ROUGE [8] for generated question efficiency.

Quantitative Results:

First, Table 1 shows the quantitative results. The results of question generation, which is the main objective of this study, are shown in the center four columns of Table 1. For both Geo and Geo+RGB, the highest accuracies were obtained when the query number was 64. Here, we found that increasing the number of query did not necessarily improve performance. We also found that the Geo+RGB (query number 64) obtained the highest performance for all four evaluation metrics for question generation. Since the proposed dataset 3D-REQ includes data that can distinguish objects by color, it can be assumed that the presence of color information is beneficial for question generation.

The right side of Table 1 shows the results of 3D object detection. Here, Geo, which uses only geometric information, achieved comparatively higher accuracy than Geo+RGB, which also uses color information. Although the proposed dataset 3D-REQ includes questions about color, it is difficult to improve accuracy by simply adding RGB because the detector used is designed to detect from geometric information. We plan to study the use of RGB values as one of our future works. Similar to that of question generation, the highest accuracies were obtained when the number of queries was 64 for both Geo and Geo+RGB.

Qualitative Results:



Object Class: Wall
Referring Expression: Touch the wall.
Question(GT): What is the color?
Questions(Predict): What is the color?

Object Class: Picture
Referring Expression: See the painted picture.
Question(GT): What is the shape?
Questions(Predict): What is the size?

Figure 2. Qualitative results of Geo+RGB (query number 64). Predicted boxes are marked with blue and ground truth boxes are marked with green. We show examples that our method produced good predictions of boxes and question (blue block) as well as a failure case (orange block).

Two example results of Geo+RGB (query number 64) are shown in Figure 2. First, on the left side, the referring expression is "Touch the wall". Here, there are a total of three walls, and our proposed method was able to detect all three walls. The color of the three walls is different, therefore the color information can be used to distinguish the walls. Here, our proposed method was also able to generate the question "What is the color?". This example shows that our proposed method is able to detect the objects pointed by the referring expression and generate sentences to distinguish them.

Another example is shown on the right. Here, the referring expression is "See the painted picture". Here, there are two small paintings. As shown in the green bounding box, each of them has a different shape. In this example, our method did not detect the painting correctly and could not generate a sentence that distinguishes between the two paintings. Currently, our proposed method still has room for improvement for small objects, and we will address this issue in the future work.

4. Conclusion

In this paper, we propose a new task and dataset aiming at eliminating the ambiguity of referring expressions in 3D indoor environments by generating questions. Our experimental results show that our proposed method can generate questions for eliminating ambiguity, and our proposed dataset can be used as a benchmark dataset for future works on disambiguation in 3D visual grounding.

References

[1] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020. 1

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *16th European Conference on Computer Vision (ECCV)*, 2020. 1

[3] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[4] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2906–2917, October 2021. 2

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 311–318, 2002. 2

[6] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2

[7] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 2

[8] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 2