Solving Zero-Shot 3D Visual Grounding as Constraint Satisfaction Problems

Qihao Yuan [1], Kailai Li [1,*], and Jiaming Zhang [2,*]

[1]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen
[2]Computer Vision for Human-Computer Interaction Lab (cv:hci), Karlsruhe Institute of Technology

*Corresponding authors with equal contribution

BMVC 2025

Abstract

3D visual grounding (3DVG) aims to locate objects in a 3D scene with natural language descriptions. Supervised methods have achieved decent accuracy, but have a closed vocabulary and limited language understanding ability. Zero-shot methods mostly utilize large language models (LLMs) to handle natural language descriptions, yet suffer from slow inference speed. To address these problems, in this work, we propose a zero-shot method that reformulates the 3DVG task as a Constraint Satisfaction Problem (CSP), where the variables and constraints represent objects and their spatial relations, respectively. This allows a global reasoning of all relevant objects, producing grounding results of both the target and anchor objects. Moreover, we demonstrate the flexibility of our framework by handling negation- and counting-based queries with only minor extra coding efforts. Our system, Constraint Satisfaction Visual Grounding (CSVG), has been extensively evaluated on the public datasets ScanRefer and Nr3D datasets using only open-source LLMs. Results show the effectiveness of CSVG and superior grounding accuracy over current state-of-the-art zero-shot 3DVG methods with improvements of $+7.0\%$ (Acc@0.5 score) and $+11.2\%$ on the ScanRefer and Nr3D datasets, respectively. The code of our system will be made publicly available upon acceptance of the work.

Background

Given a query involving multiple object relationships, existing 3D visual grounding method like ZSVG3D reasons about the individual object relationship locally one at a time, which lacks a global perspective all relationships as a whole. This often results in incorrect grounding given non-trivial queries in complex 3D scenes (Fig.). In this paper, we introduce a novel zero-shot 3D visual grounding framework, Constraint Satisfaction Visual Grounding (CSVG), by enabling global interpretation of object spatial relations through the formulation of Constraint Satisfaction Problem (CSP).

Anchor Object Grounding

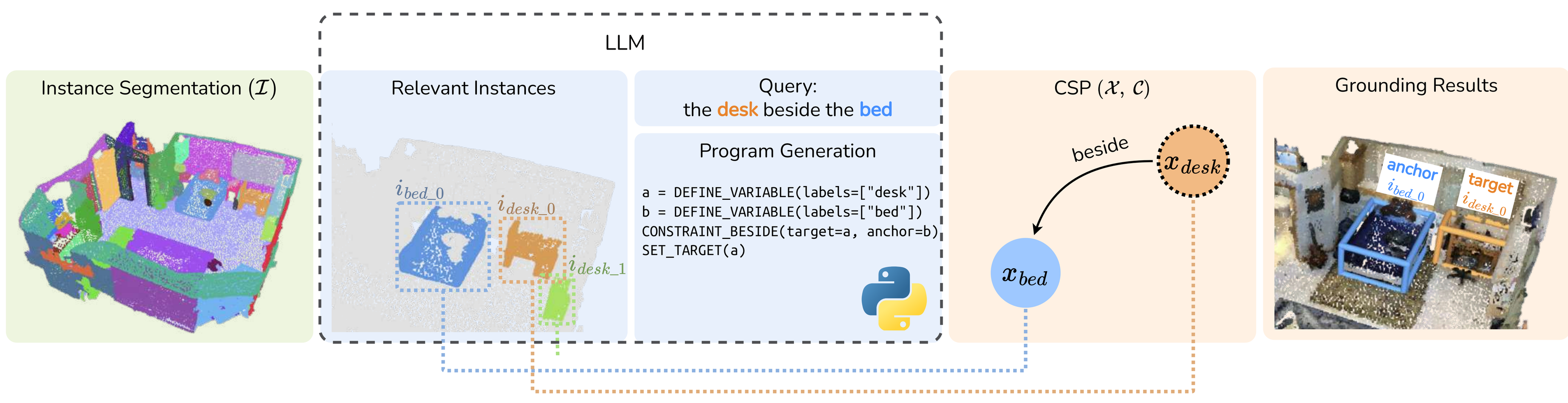
Contributions

- We formulate the 3DVG task as a constraint satisfaction problem, which allows for global reasoning about spatial relations of all relevant objects at once, leading to grounding of both the target and anchor objects.
- We demonstrate the flexibility of our framework by extending it to handle negation- and counting-based queries with only minor code modifications. In contrast, supervised methods would require a substantial amount of additional training data for theses capabilities.
- We evaluate our system on the well-known Nr3D and ScanRefer datasets, achieving increases in accuracy (Acc@0.5 score) over the state of the art (ZSVG3D) by a margin of $+$**11.2**% on the Nr3D and $+$**7.0**% on the ScanRefer dataset.

System Design

Our system works in this way:

1. first segment the input point cloud with Mask3D (a supervised instance segmentation model for point cloud).
2. Instances with the same label will be grouped into candidate objects for that label.
3. Then we prompt an LLM to formulate the CSP problem based on the language query.
4. A CSP solver then selects for each label a candidate object.
5. Finally, the candidate object for the target label is the grounding result.



System Pipeline

As an example, here we have the query: *the desk beside the bed.* This query is given to the LLM to generate a CSP instance:

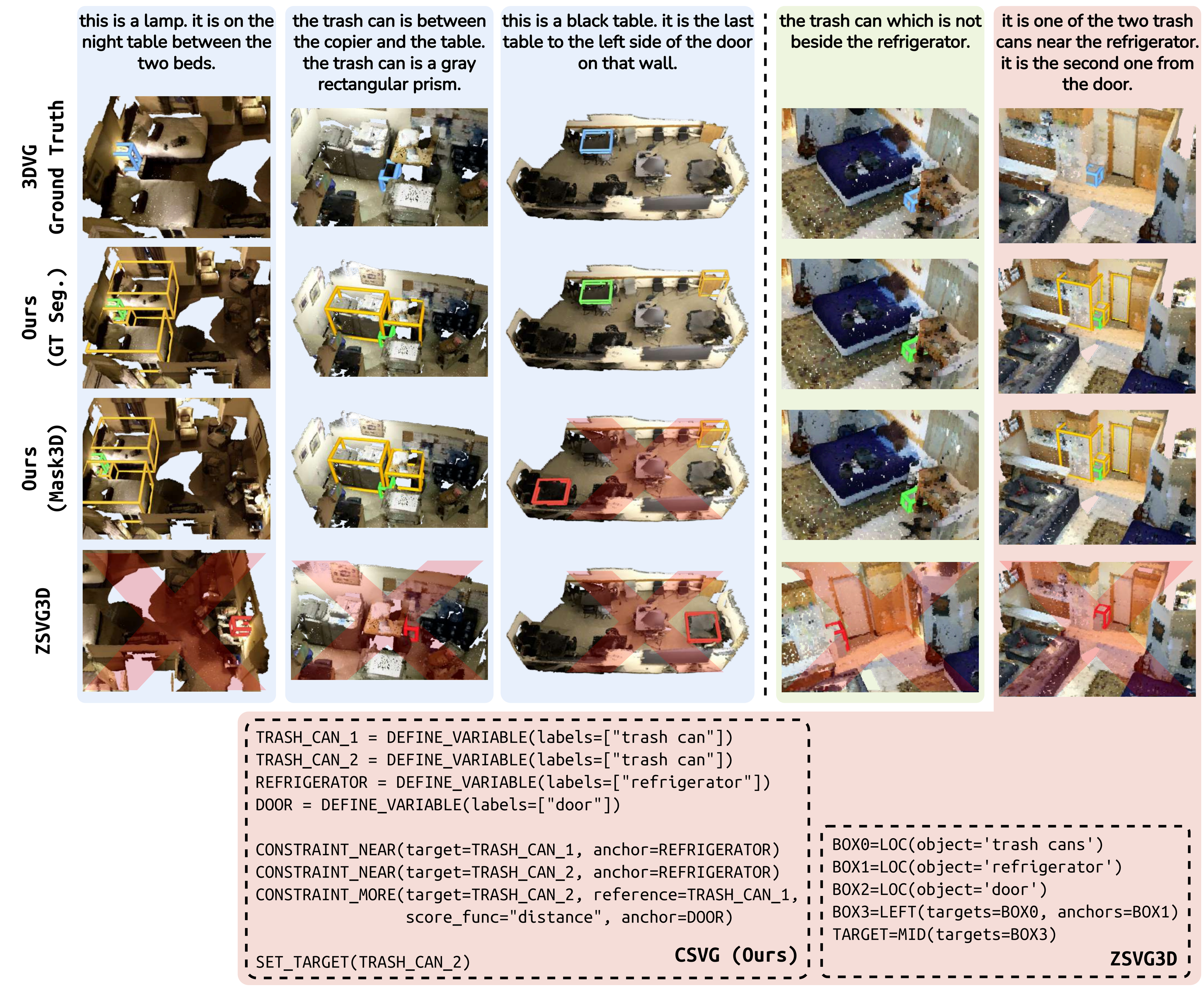
```
a = DEFINE_VARIABLE(labels=["desk"])
b = DEFINE_VARIABLE(labels=["bed"])
CONSTRAINT_BESIDE(target=a, anchor=b)
SET_TARGET(a)
```

In the point cloud, we use Mask3D to segment a bed (i_{bed_0}), and two desks (i_{desk_0}, i_{desk_1}). Therefore, the label "bed" has one candidate, and the label "desk" has two candidate objects. We then run the CSP solver to find an assigment from candidates to labels that satisfies the constraints. In the end, we get these assignments:

$$bed \leftarrow i_{bed\_0} \qquad desk \leftarrow i_{desk\_0}$$

The LLM identifies the desk as the target (via `SET_TARGET(a)`). Therefore, the segmented object, i_{desk_0} is the grounding target, whose bounding has already been obtained from Mask3D.

Evaluation & Conclusion



Find our open-source implementation:

Acknowledgement: We would like to thank Computer Vision for Human-Computer Interaction Lab (cv:hci), led by Prof. Dr.-Ing. Rainer Stiefelhagen, at Karlsruhe Institute of Technology, for providing GPUs and other support to our work.

university of groningen

ASIG Agile Sensing & Intelligence Group

KIT Karlsruhe Institute of Technology