

# Instance-Level Semantic Maps for Vision Language Navigation

Laksh Nanwani<sup>1</sup>, Anmol Agarwal<sup>1</sup>, Kanishk Jain<sup>1</sup>, Raghav Prabhakar<sup>1</sup>,  
Aaron Monis<sup>1</sup>, Aditya Mathur<sup>1</sup>, Krishna Murthy Jatavallabhula<sup>2</sup>,  
A. H. Abdul Hafez<sup>3</sup>, Vineet Gandhi<sup>1</sup>, K. Madhava Krishna<sup>1</sup>

Query: Walk to the 4th chair in your field of view

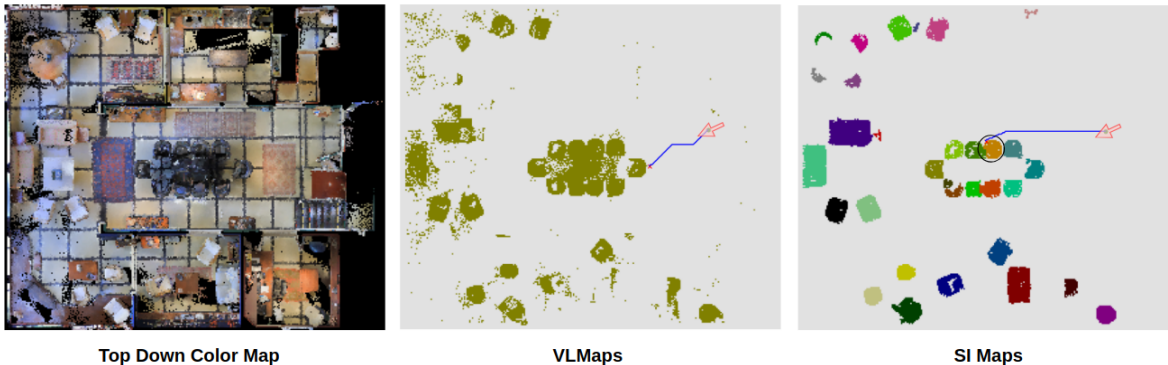


Fig. 1: VLMaps [1], being a semantic top-view map, cannot distinguish between different instances of the same object. On the other hand, SI Maps (ours) are directly amenable for handling such queries as they contain instance-specific information for all objects in the environment. For the scene on the extreme left, the instances of the object ‘chair’ as detected by SI Maps is shown in different colors in the rightmost figure.

**Abstract**—Humans have a natural ability to perform semantic associations with the surrounding objects in the environment. This allows them to create a mental map of the environment, allowing them to navigate on-demand when given linguistic instructions. A natural goal in Vision Language Navigation (VLN) research is to impart autonomous agents with similar capabilities. Recent works take a step towards this goal by creating a semantic spatial map representation of the environment without any labeled data. However, their representations are limited for practical applicability as they do not distinguish between different instances of the same object. In this work, we address this limitation by integrating instance-level information into spatial map representation using a community detection algorithm and utilizing word ontology learned by large language models (LLMs) to perform open-set semantic associations in the mapping representation. The resulting map representation improves the navigation performance by two-fold (233%) on realistic language commands with instance-specific descriptions compared to the baseline. We validate the practicality and effectiveness of our approach through extensive qualitative and quantitative experiments.

## I. INTRODUCTION

Advancements in machine learning research have brought about rapid changes in the field of robotics, allowing for the development of sophisticated autonomous agents. However, making this technology practically viable for large-scale adoption requires a natural mechanism to interact with humans. Vision Language Navigation (VLN) research aims

to achieve this goal by incorporating natural language understanding into autonomous agents to navigate the environment based on linguistic commands. Prior approaches to VLN have addressed this task by harnessing the capabilities of visual grounding models, which allow the navigating agents to localize objects in the visual scene or directly ground navigable regions based on linguistic descriptions. However, these approaches fail to address linguistic commands which require spatial precision to identify the goal region. Furthermore, these approaches assume that the object referred to by the linguistic command is always visible in the current scene. Such an assumption rarely holds in realistic scenarios, where things can move in or out of the current scene as we navigate the environment.

Consider the example in Figure 1 with the language command, “walk to the fourth chair in your field of view”. To execute this command, we first need to explore the entire room to find all instances of chairs and then find the fourth instance from where the command was given. For visual grounding-based approaches, it is non-trivial to handle such scenarios as there is no way to rank the localized chairs based on distance. To counteract the above issues, geometric maps, which create a global mapping of the surrounding environment, provide a direct mechanism to ground all the objects present in the scene, including those not visible in the current view, and additionally, are readily amenable for planning and navigation purposes. In this work, we propose a memory-efficient mechanism for creating a semantic spatial representation of the environment, which is directly applica-

<sup>1</sup>KCIS, International Institute of Information Technology, Hyderabad, India.

<sup>2</sup>CSAIL, MIT, Cambridge, United States.

<sup>3</sup>Hasan Kalyoncu University, Sahinbey, Gaziantep, Turkey.

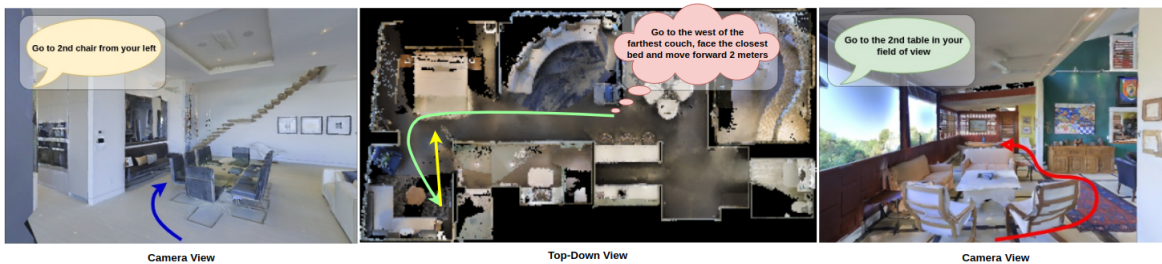


Fig. 2: Our top-view map representation allows indoor embodied agents to perform complex instance-specific goal navigation in object-rich environments. The language queries can refer to individual instances based on spatial and viewpoint configuration with respect to other objects of the same type while preserving the navigation performance on standard language queries.

ble to robots navigating in real-world scenes.

Recent works like VLMaps and NLMap [2] propose a mechanism to build semantic spatial maps without any labeled data by fusing pre-trained vision-language features with the 3D point cloud of the physical world. They compute the similarity between visual and linguistic features in a common semantic space of a large-scale pre-trained vision-language model and utilize large-language models to convert the natural language command to a sequence of navigation goals for planning. However, their map representation doesn't allow them to differentiate between different instances of the same object and hence handle language queries that describe an instance-specific navigation goal, like the ones mentioned in Figure 2, as the visual encodings are instance-agnostic. Moreover, their mechanism is memory intensive as they require high-dimensional feature embeddings to make semantic associations for the objects in the visual scene.

Our work focuses on creating spatial maps of the environment with instance-level semantics. We achieve this in a memory-efficient manner, bypassing the use of feature embeddings altogether. We show that **Semantic Instance Maps (SI Maps)** are computationally efficient to construct and allow for a wide range of complex and realistic commands that evade prior works.

## II. RELATED WORK

### A. Semantic Mapping

With the recent progress in computer vision and natural language processing literature, there has been considerable interest in augmenting the semantic understanding of traditional SLAM algorithms. Earlier works like SLAM++ [3] propose an object-oriented SLAM, which utilizes prior knowledge about the domain-specific objects and structures in the environment. Later works like [4] assign instance-level semantics using Mask-RCNN [5] to 3D volumetric maps. Some methods [1], [2] have also explored transferring predictions from CNNs in 2D pixel space to 3D space for 3D reconstruction. Concurrent to our work, [6] proposes a deep reinforcement learning-based approach for multi-object instance navigation, albeit without linguistic commands. VLMaps [1] and NLMap-Saycan [2] propose a natural language queryable scene representation with Visual Language models (VLMs). These methods utilize large-language

models (LLMs) to parse language instructions and identify the involved objects to query the scene representation for object availability and location.

### B. Instance Segmentation

The ability to identify and localize different instances of similar objects is crucial for visual perception tasks in robotics. In the Computer Vision literature, the task of instance segmentation serves to evaluate such capabilities formally. Earlier works [5] utilized region proposal networks to predict candidate bounding boxes followed by a mask head to regress the instance-level segmentation mask for each proposal. While initial approaches designed task-specific architectures, more recent methods [7] have moved towards generalized architectures for different image segmentation tasks like semantic, instance, and panoptic segmentation. Mask2Former [7] employs attention mechanism to extract localized object-centric features in an end-end manner. In this work, we utilize segmentation masks from Mask2Former to create instance-level semantic maps which are directly amenable for planning during autonomous navigation.

### C. Vision Language Navigation

Most of the work in Vision Language Navigation (VLN) has focused on navigating in the environment using semantic perception based on the front camera view of the autonomous agent. Specifically, these works take the front camera image and the language command as input, and the navigation task is reduced to a sequence modeling task where at each time stamp, the optimal action is predicted to complete the navigation task successfully. Subsequent works have tackled the VLN problem using sequence-to-sequence learning [8], reinforcement learning [9] or behavior cloning methods [10]. However, these methods are non-trivial to interpret, and recent works [8] have found that such methods are unable to utilize the visual modality effectively for the navigation task. Consequently, recent works [1], [2] on VLN have focused on creating a semantic map of the environment for motion planning and utilizing visual grounding capabilities of large-scale vision-language models [11] to ground the semantic concepts in a visual world. In this work, we focus on creating a semantic mapping representation of the environment using large-scale language models. Unlike prior works, we create these maps in an embedding-free manner, thus reducing the computational cost significantly.

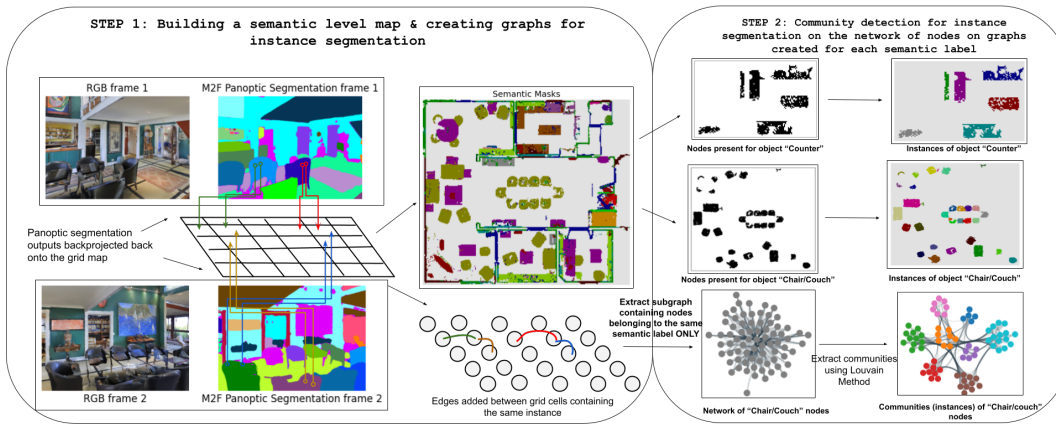


Fig. 3: In STEP 1, we create a semantic level map of the environment by back projecting the Mask2Former semantic labels of the RGB pixels across different images onto the grid map. In STEP 2, we extract the subgraph concerned with object  $o$  and run a community detection algorithm to break the grid cells containing object  $o$  into instances.

### III. METHOD

#### A. Problem Statement

In this work, we aim to create a semantic map of the surrounding environment containing instance-level information for the various objects. Maps containing both instance-level and semantic information are necessary to handle linguistic commands which are frequently used in the daily vernacular. For example, consider the command, “Go to the empty chair near the third table”. We are required to identify “which instance of the table” is being talked about and then point out the instance of the empty chair. Our approach is equipped to handle such scenarios through an instance-specific mapping representation of the environment. We build SI Maps using only RGB-D sensors, pose information, and an off-the-shelf panoptic segmentation model. SI Maps creation involves two steps: (1) Occupancy map creation with semantic labels and (2) Community detection to separate instances of a given semantic label. The whole pipeline is illustrated in Figure 3.

#### B. SI Map Creation

**Building Occupancy Grid:** We define SI Maps as  $\mathcal{M} \in \mathbb{R}^{\bar{H} \times \bar{W} \times 2}$ , where  $\bar{H}$  and  $\bar{W}$  represent the size of the top-down grid map. Similar to VLMaps, with the scale parameter  $s$  ( $= 0.05m$  in our experiments), a SI Map  $\mathcal{M}$  represents an area with size  $s\bar{H} \times s\bar{W}$  square meters.  $\mathcal{M}_{i,j} = \langle o, t \rangle$  means that grid cell  $(i, j)$  is occupied by the  $t^{\text{th}}$  instance of object  $o$  in the environment. Since we are using the Mask2Former panoptic segmentation model trained on the COCO dataset [12],  $o \in \mathcal{O}$  (where  $\mathcal{O}$  is the set of objects present in the COCO dataset). To build our map, similar to VLMaps, we, for each RGB-D frame, back-project all the depth pixels  $\mathbf{u} = (u, v)$  to form a local depth point cloud that we transform to the world frame using the pose information. For depth pixel  $\mathbf{u} = (u, v)$  belonging to the  $i^{\text{th}}$  RGB-D frame, let  $(p_x^{i,u,v}, p_y^{i,u,v})$  represent the coordinates of the projected point in the grid map  $\mathcal{M}$ .

**Integrating Instance-level information:** With the occupancy map defined, we now utilize community detection algorithms to separate out the different instances in the

environment. Specifically, we use the modularity-based Louvain method, a greedy, hierarchical optimization method that iteratively refines communities to maximize the modularity value. The modularity value is a measure of the density of links within communities compared to links between communities.

Let the output of the panoptic segmentation model for  $\mathbf{u} = (u, v)$  be  $\langle o_{i,u,v}, t_{i,u,v} \rangle$ . This means object  $o_{i,u,v}$ 's  $t_{i,u,v}^{\text{th}}$  instance within the frame is present at pixel  $\mathbf{u}$ . We use this information to set the object label  $o$  for  $\mathcal{M}_{(p_x^{i,u,v}, p_y^{i,u,v})}$  as  $o_{i,u,v}$ . When there exist multiple 3D depth pixels projecting to the same grid location in the map, we retain the label of the pixel with the highest vertical height.

To divide the different grid cells labeled having object  $o$  into different instances, we construct an undirected weighted graph  $G = (V, E, W)$ , where each grid cell  $(i, j)$  for whom the object label of  $\mathcal{M}_{i,j}$  is equal to  $o$  is included as a node in the set of vertices  $V$ . Whenever two neighbouring pixels  $\mathbf{u}_1 = (u_1, v_1)$  and  $\mathbf{u}_2 = (u_2, v_2)$  belong to the same entity in the  $i^{\text{th}}$  RGB-D frame, their corresponding grid cells  $(p_x^{i,u_1,v_1}, p_y^{i,u_1,v_1})$  and  $(p_x^{i,u_2,v_2}, p_y^{i,u_2,v_2})$  should also belong to the same instance in real-world. Hence, whenever pixels  $\mathbf{u}_1$  and  $\mathbf{u}_2$  have the semantic label  $o$  and  $\langle o_{i,u_1,v_1}, t_{i,u_1,v_1} \rangle = \langle o_{i,u_2,v_2}, t_{i,u_2,v_2} \rangle$ , i.e., depth pixels  $\mathbf{u}_1$  and  $\mathbf{u}_2$  belong to the same entity within the image, we increase the edge weight between grid cells  $(p_x^{i,u_1,v_1}, p_y^{i,u_1,v_1})$  and  $(p_x^{i,u_2,v_2}, p_y^{i,u_2,v_2})$  by one. This helps us in transferring the instance segmentation information present in the panoptic segmentation outputs of the RGB-D frames to our map and also helps us to track the same instance across frames using the pose data. To prevent the frequency of visiting a particular area in the environment during mapping from unfairly affecting any edge weight, we normalize all the edge weights by the number of times their constituent nodes (grid cells) were observed across all RGB-D images for that scene. Ideally, in our graph, all grid cells belonging to the same connected component should belong to the same real-world entity. But Mask2Former masks are not perfect at a pixel level; hence it is possible for spurious edges to be drawn between nodes belonging to different real-world

entities. However, such edges are likely to be few in number. To disregard such spurious edges, we group the nodes in  $V$  using community detection algorithms instead of naively breaking them into connected components.

We initialize the graph with a separate community for each node. We use the Louvain community detection method, which involves two phases: (1) Modularity optimization and (2) Community aggregation. During modularity optimization, for each node in the graph, we compute the change in modularity by moving it to neighboring communities. The node is transferred to the community, which results in the highest increase in modularity. This procedure is repeated for all nodes until no further improvement in modularity is possible. In the community aggregation phase, the communities formed in the modularity optimization phase are considered single nodes. The weights of the edges between the new nodes are determined by the sum of the weights of the edges between the nodes in the original communities. The two phases are iteratively repeated until the modularity value converges. After convergence, we get a labeled graph, where the nodes are grouped based on their community membership, i.e., occupancy grid cells belonging to the same instance are grouped together for all the objects in the environment. To correct the over-segmentation of communities, a post-processing step is applied to merge communities  $C_1$  and  $C_2$  if more than  $K\%$  of the members of  $C_1$  are neighbors of some member of  $C_2$ .

In contrast to VLMaps, our approach doesn't utilize the high dimensional LSeg [13] feature embeddings for semantic map creation, which provides a memory-efficient mechanism to construct the instance-level semantic occupancy grid. For comparison, VLMaps representation requires an average storage of about 2 gigabytes for a  $1000 \times 1000$  map, whereas SI Maps needs only about 16 megabytes for the same map size. Additionally, the proposed approach is highly flexible and adaptable, as it can easily incorporate other types of sensor data like LiDar, IMU and plug different segmentation models. The provision of tunable hyper-parameter  $K$  further provides controllability in our approach, which is a desired capability for real-world deployment. In the next section, we show how SI Maps can be directly used for language-conditioned navigation.

### C. Language-based Navigation

The significance of Semantic Instance maps becomes apparent when dealing with commands that necessitate instance-level grounding. For a given language command, we would like to identify the region in SI Maps where the robot must navigate to execute the command successfully. Additionally, since different commands can refer to different navigational maneuvers, we must also determine the maneuvers required for a specific language query. To achieve this, we define function primitives for each possible maneuver, reducing the task to classifying the appropriate function primitive for each sub-command. For this classification, we utilize the powerful large language model (LLM), ChatGPT[14], for motion planning.

```
# navigate to fourth closest chair

chair = bot.get_closest_instance('chair', instance=4)
if chair is None:
    exit(0)
bot.navigate(chair)

# walk to the third cabinet to your left and then
walk to the 2nd chair closest to it

cabinet = bot.to_your_left('cabinet', instance=3)
bot.navigate(cabinet)
chair = bot.get_closest_instance('chair', instance=2)
if chair is None:
    exit(0)
bot.navigate(chair)
```

Fig. 4: An example of the executable Python code generated by ChatGPT for the given language commands. The generated code includes an instance parameter in the function primitive call for navigating to the specified instance in the environment.

LLMs, trained on billions of lines of text and code, demonstrate advanced natural language understanding, reasoning, and coding capabilities. Similar to the approach with VLMaps, we repurpose LLMs to generate executable Python code for the robot. Specifically, we supply ChatGPT with the list of function primitives and their respective descriptions. We then prompt ChatGPT with several language queries accompanied by the corresponding ground truth Python code containing a sequence of function primitives based on the language command. During inference, for each language command, we provide ChatGPT with the list of objects present in the SI Maps and generate Python code that refers to the specific instances involved in the language command.

In Figure 4, we show a few examples of the Python executable code generated by ChatGPT for the given commands. ChatGPT successfully generates the correct executable code after prompting it with a few examples of language queries and corresponding ground truth Python executable code. To ground instances, our function primitives calls also include an instance parameter to handle instance-specific queries. The instance parameter is directly inferred from the language command by ChatGPT along with the object of interest. Overall, we define 23 function primitives for complex navigational maneuvers like moving between two objects, navigating to  $n^{th}$  closest object, etc., and the essential turning and moving primitives.

## IV. EXPERIMENTS

### A. Experimental Setup

We showcase the effectiveness of our approach on multiple scenes from Matterport3D [15] dataset in the Habitat [16] simulator. Matterport3D is a commonly used dataset for evaluating the navigational capabilities of existing VLN agents in an indoor environment. The robot must maneuver in a continuous environment, performing navigational maneuvers

specified by the natural language command. For top-view map creation, we collect 5,267 RGB-D frames from 5 different scenes and store the camera pose for each frame.

**Baseline:** We evaluate against a logical baseline where the semantic top-view maps from the VLMaps-based approach are separated into separate instances. If the objects in the environment are well separated, the semantic segmentation output should already contain the information required to separate different instances of similar objects by simply applying connected components. As a result, our baseline involves applying connected components over the VLMaps output. However, in realistic scenarios, different instances of the same object can be close to each other; for example: in a restaurant, chairs belonging to the same table are close to each other. In such a scenario, just computing connected components will not work, as multiple instances will get clubbed into a single instance.

**Evaluation Metrics:** Like prior approaches [1], [8], [17] in VLN literature, we use the gold standard *Success Rate* metric, also known as *Task Completion* metric to measure the success ratio for the navigation task. We compute the *Success Rate* metric through human and automatic evaluations. For automatic evaluation, we use the ground truth environment map and compute the *Success Rate* using a pre-defined heuristic where the navigation sub-goal is considered successful if we stop within a threshold distance of the ground truth object. For human evaluation, we verify if the agent ends up in a position desired according to the query.

### B. Evaluation Results

In this section, we perform quantitative and qualitative comparisons of SI Maps against VLMaps and VLMaps with connected components. We compare the performance of each scene representation for the downstream language-based navigation task using the *Success Rate* in table I. We use the same function primitives for all the methods.

Human evaluation was done because of the observation made during a few queries where the agent ended up close to the target object, but it did not complete the task in the desired way.

Method	Success Rate	
	Human Evaluation	Automatic Evaluation
VL Maps	0.24	0.46
VL Maps with CC	0.34	0.48
SI Maps (K=5)	<b>0.80</b>	<b>0.88</b>
SI Maps (K=9)	0.76	<b>0.88</b>

TABLE I: SI Maps outperform other baseline methods by significantly large margins on the *Success Rate* metric. The best results are highlighted in **bold**.

We observe that SI Maps exhibit a remarkable improvement in performance compared to other approaches. SI Maps achieve an impressive two-fold increase in success rate metric compared to 24% obtained by VLMaps on human evaluation, demonstrating a substantial leap in the instance-specific goal navigation. Since VLMaps only contain seman-

tic information, they fail on queries that refer to specific instances of an object, like “navigate to the second counter”. Our logical baseline, VLMaps with connected components, can handle some instance-specific queries, resulting in an incremental performance gain of 10% for human evaluation than vanilla VLMaps. However, the success of this method is observed in scenes where neighboring instances of the same object have ample room between them. In contrast, real-life environments such as offices, restaurants, and hospitals often have objects in close proximity to each other. In these cases, instance-level information is essential for distinguishing between neighboring objects. SI Maps demonstrate robustness to object placement in the environment by directly utilizing the instance-level information provided by the instance segmentation model during the occupancy grid creation.

### C. Qualitative Results

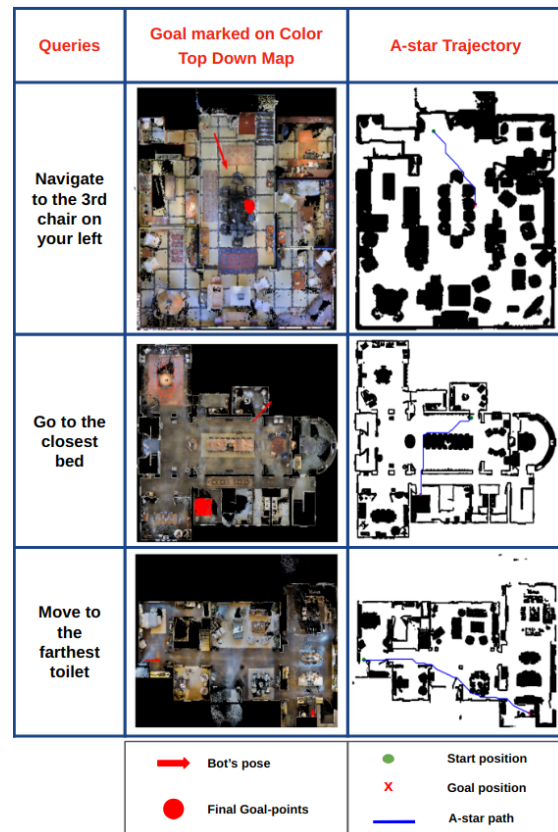


Fig. 5: The above figure shows the agent in different scenes in a simulated environment with three different queries. Images on the top show the RGB top-down view map, along with the segmented goal object instance. The corresponding images on the bottom represent the path taken by the agent to reach the desired object from the initial location.

In this section, we showcase qualitative examples of our approach for the vision language navigation task. The results are illustrated in Figure 5 with the corresponding A-star trajectory using SI Maps for navigation. SI Maps allow navigating to specific instances in the scene based on their relative distance with respect to other objects (left, center)

and direction-based specification in the global map (right). The downstream navigation, as a consequence of SI Maps, is agnostic to the starting pose and orientation of the agent in the environment.

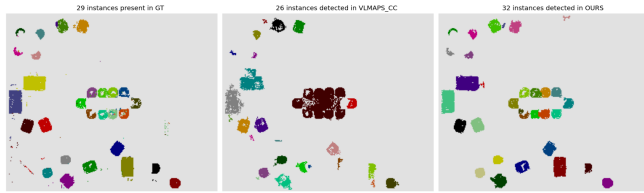


Fig. 6: Qualitative example of the instance-level semantics captured by different methods for all the chairs in the environment. SI Maps clearly localize the different instances in the map.

We also show qualitative comparisons of different methods on the quality of instance-level top-view maps in Figure 6 for different seating objects (chair, couch, sofa) in the simulated environment. Our approach effectively captures the instance-level semantics of objects in the environment, recovering 32 instances out of 29 present in the map (with 3 extra noisy segments). In contrast, the baseline of VLMaps with connected components detects 26 instances, but most of them are noisy segments, and it merges several separate instances (for the same object in close proximity) into a single instance. Our results are particularly impressive in the middle region of the map, which corresponds to the dining area in the environment. Here, the chairs are in close proximity to each other, and the vanilla VLMaps approach fails when a particular instance of chair is queried. Similarly, applying connected components-based heuristics to separate instances is not enough, as the semantic segmentation masks of the chairs end up being connected with each other, resulting in multiple instances being merged.

The VLMaps-based approaches rely on alignment between per-pixel visual embeddings and linguistic feature embeddings, which can be sensitive to noise due to the unconstrained nature of the association. The benefit of our feature-embedding-free approach becomes evident as we directly constrain the occupancy grid creation with the instance segmentation masks. As a result, SI Maps have considerably less noise than derivative VLMaps approaches. Community detection further helps reduce noise by filtering out spurious communities formed due to noise, leading to a much cleaner map, which can also be observed in Figures 1, 6.

## V. CONCLUSION

In this study, we introduce a novel instance-focused scene representation for indoor settings, enabling seamless language-based navigation across various environments. Our representation accommodates language commands that refer to specific instances within the environment. Furthermore, our map creation method is more memory-efficient, resulting in an impressive 128-fold decrease in storage, as it does not rely on high-dimensional feature embeddings for visual and linguistic modalities. Additionally, our approach

demonstrates robustness in relation to object placement in the environment and is less vulnerable to noise than previous methods. We showcase the practicality of the proposed SI Maps using success rate and panoptic quality metrics. Future research could investigate 3D instance segmentation techniques to incorporate instance-level semantics into the occupancy grid creation process directly.

## ACKNOWLEDGEMENT

We acknowledge iHub-Data IIIT Hyderabad for their support to this work.

## REFERENCES

- [1] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, (London, UK), 2023.
- [2] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler, "Open-vocabulary queryable scene representations for real world planning," in *arXiv preprint arXiv:2209.09874*, 2022.
- [3] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1352–1359, 2013.
- [4] J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger, "Fusion++: Volumetric object-level slam," in *2018 international conference on 3D vision (3DV)*, pp. 32–41, IEEE, 2018.
- [5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- [6] N. Gireesh, A. Agrawal, A. Datta, S. Banerjee, M. Sridharan, B. Bhowmick, and M. Krishna, "Sequence-agnostic multi-object navigation," in *IEEE International Conference on Robotics and Automation*, 2023. (to be published).
- [7] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," 2022.
- [8] R. Schumann and S. Riezler, "Analyzing generalization of vision and language navigation to unseen outdoor areas," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Dublin, Ireland), pp. 7519–7532, Association for Computational Linguistics, May 2022.
- [9] K. Nguyen, D. Dey, C. Brockett, and B. Dolan, "Vision-based navigation with language-based assistance via imitation learning with indirect intervention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12527–12537, 2019.
- [10] A. Das, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Neural modular control for embodied question answering," in *Conference on Robot Learning*, pp. 53–62, PMLR, 2018.
- [11] 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*, pp. 8748–8763, PMLR, 2021.
- [12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.
- [13] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *International Conference on Learning Representations*, 2022.
- [14] OpenAI, "Chatgpt." <https://openai.com/blog/chatgpt>.
- [15] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.
- [16] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9339–9347, 2019.
- [17] K. Jain, V. Chhangani, A. Tiwari, K. M. Krishna, and V. Gandhi, "Ground then navigate: Language-guided navigation in dynamic scenes," *arXiv preprint arXiv:2209.11972*, 2022.