Multi-Modal Open-Vocabulary 3D Scene Understanding for Indoor environments

Arun Madhusudhanan
madhusudhanan.a@northeastern.edu

Shankara Narayanan Vaidyanathan
vaidyanathan.sh@northeastern.edu

Tejaswini Dilip Deore
deore.t@northeastern.edu

Abstract

3D scene understanding is a specific computer vision task aimed at understanding the semantics, physical properties, functions of objects in a given 3D representation of the world (like point clouds or meshes). Traditionally, scene understanding tasks like semantic segmentation have been limited to a closed set of objects. Recent approaches have addressed this by co-embedding features of 3D points with text and image pixels in the CLIP feature space. This makes them fundamentally open-set. This allows zero-shot scene understanding.

In our work, we explore extensions to such semantically rich 3D representations to assist with robot perception and planning. Towards this, we explore hierarchical representations of 3D scenes, comprising semantically annotated point clouds, objects and rooms. By embedding our 3D scene representation in the CLIP feature space, we further expand the capability to incorporate additional modalities like audio, alongside text and image inputs. We perform qualitative and quantitative evaluations to drive the need for hierarchical representations and multi modal capabilities to perform scene understanding.

1. Introduction

Robots require a 3D representation of a scene to interact with the environment. For this to be efficient, such representations shouldn't be limited to a fixed set of objects; instead, it must be open-vocabulary. In today's robots we provide geometric commands like "grasp object at Pose X". In contrast, we humans are able to handle higher level understanding like "grasp toy in the living room" which involves understanding spatial context, object categories, and environmental semantics. To achieve this, robots need higher-level topological information and reasoning beyond simple geometric positioning. In addition, making these maps multi-modal would allow us to perform additional tasks such as object retrieval using modalities like text queries, images, audio etc. Thus, we need the map representation to be hierarchical, open-set and multi-modal.

In recent times, foundation models have shown remarkable capabilities to perform a wide variety of open vocabulary challenges in 2D vision and text. These models are trained on an internet scale of data to achieve significant generalization. They don't need additional re-training or fine-tuning. For instance, image-only foundation models like DINO [7] and SAM [5] excel in classification, detection, and segmentation, while image-language models like CLIP excel in learning robust text-image representations. However, unlike images or text, we don't have such internet scale data to build such foundation models for such 3D related aspects. Recent approaches have addressed this by utilizing foundation models of other modalities and mapping 3D point representations to their feature spaces. This has shown great promise to perform tasks like zero-shot segmentation, affordance estimation and 3D object retrieval.

Another aspect is that current works often lack contextual understanding. For example, searching for "a pillow in the bedroom" highlights pillows in all rooms since there is no higher level topological understanding. Towards this, we explore leveraging 3D scene graphs. Using such a graphical structure helps us encode 3D points, objects, and rooms as nodes, with edges representing relationships between points-objects and objects-rooms. This provides various levels of abstraction.

Additionally, we can lift multiple modalities into a shared feature space, like in CLIP. By integrating backbones such as AudioCLIP [3] and ImageBind [1], we can create inherently multi-modal maps that support queries across different modalities—text, images, and audio.

Thus, in this work, we build a simple pipeline to build a map to help us perform efficient scene understanding. We make use of foundation models to expand the 3D setting to an open vocabulary setting, make the maps multi-modal and use scene graphs to build a hierarchical representation.

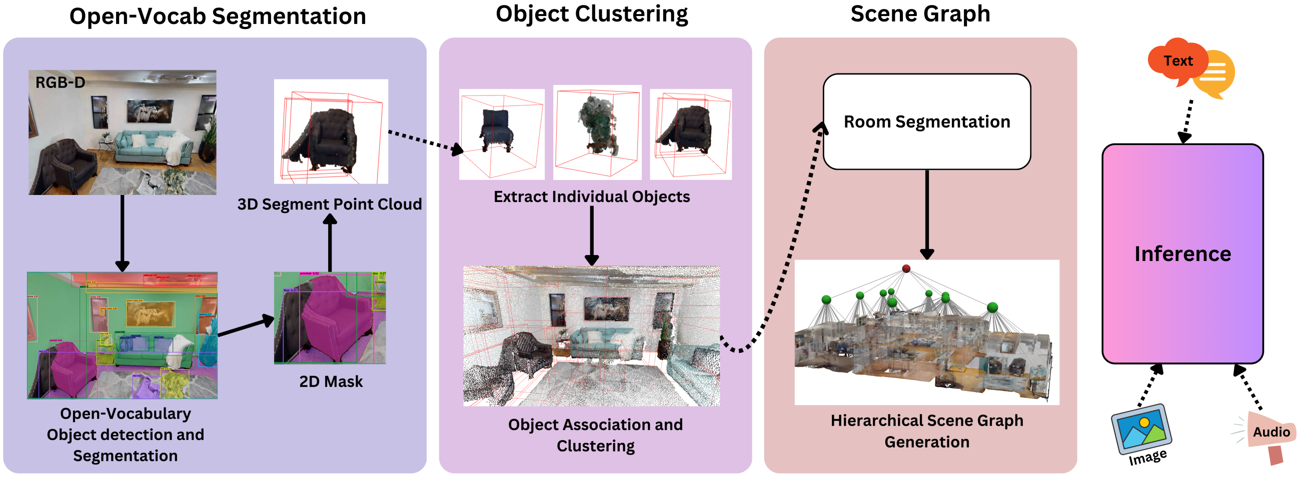

Open-Vocab Segmentation
Object Clustering
Scene Graph
RGB-D
3D Segment Point Cloud
Extract Individual Objects
Room Segmentation
Text
Inference
Open-Vocabulary Object detection and Segmentation
2D Mask
Object Association and Clustering
Hierarchical Scene Graph Generation
Image
Audio

Figure 1. Inspired by recent works, we build a simple pipeline where given a set of RGB-D images, we first perform open-vocabulary object detection and semantic segmentation to obtain 2D object masks. Leveraging the depth information, we extract corresponding 3D point clouds for each segmented object. We then associate objects across frames and cluster 3D points to create consistent object representations. We then construct a hierarchical scene graph that captures the spatial and semantic relationships within the 3D map. Multi-modal inference is performed to test reasoning across text, audio, and image modalities.

2. Relevant Works

The core idea of our project comes from OpenScene [8], which performs 3D scene understanding tasks using arbitrary text queries. They go beyond a closed set of objects by proposing a Zero-Shot Learning Approach to co-embed dense features of 3D points with image pixels and text in the CLIP feature space. However, they produce maps with per-point feature vectors. Thus, it doesn't encode inter-object relationships which is crucial for spatial reasoning. They also don't fully explore the multi-modal capabilities beyond text and image.

ConceptFusion [4] and ConceptGraphs [2] extend the OpenScene [8] work to build open vocabulary 3D maps aimed at solving robotics challenges. In ConceptFusion [4], they build a multi-modal 3D map enabling inference across additional modalities like Audio, Click etc. It lacks a scene graph representation to provide further layers of abstraction. ConceptGraphs [2] advances the approach by constructing an object-centric 3D scene graph, encoding inter-object relationships through LLM-based inference. While this provides one abstraction layer beyond point-based mapping, the method remains limited by its inability to encode broader hierarchical structures like room-level contexts and do not fully explore additional multi-modal inferencing. In HOV-SG [12], they showcase the usage of hierarchical scene graphs by building additional layers of abstraction but don't show additional multi-modal capabilities and the pipeline is significantly slower compared to the other methods.

We combine the strengths of all the above to build an open-vocabulary, multi-modal, hierarchical scene graph for efficient 3D scene understanding.

3. Method

Inspired by previous works, we build a simple pipeline that uses a posed set of RGB-D images to build a hierarchical, semantically rich representation of the environment. Figure.1 illustrates our approach.

3.1. Open-Vocab Segmentation

We first extract image tags using the Recognize Anything Model (RAM) [13] for the given series of RGB-D images, to provide a high-level understanding of the scene. GroundingDINO [6] is then used to detect object positions, and the Segment Anything Model [5] generates 2D object masks. Each extracted mask is passed to a visual feature extractor to obtain a visual descriptor. Here we explored different variants in literature.

Conceptgraph [2] directly uses OpenCLIP [9] with ViT-H-14 backbone as a visual image encoder. ConceptFusion [4] uses a weighted sum of the OpenCLIP [9] feature of the cropped image (f_l) and global CLIP feature of the full image (f_g) as a visual descriptor (f_i):

$$f_i = w_g f_g + (1 - w_g) f_l, \quad (1)$$

HOV-SG [12] uses a weighted sum of the OpenCLIP [9] features of the cropped image (f_l), cropped image without background (f_m), and global CLIP feature of the full image (f_g) as a visual descriptor (f_i):

$$f_i = w_g f_g + (1 - w_g)(w_m f_m + (1 - w_m) f_l), \quad (2)$$

Here, f_i is the final integrated feature, f_g is the global image feature, f_l is the local (cropped) image feature, and f_m is the modified local feature (with background-removed). Correspondingly, w_g is the weight for the global feature and w_m is the weight for the masked local feature.

The global feature weight w_g is determined by computing the cosine similarity between the local feature vector f_l and the global image feature f_g and applying softmax to the similarity values across local features.

$$w_g = \text{softmax}\left(\frac{f_l \cdot f_g}{||f_l|| \cdot ||f_g||}\right) \tag{3}$$

After analysis, we adopted HOV-SG's [12] feature extraction method, with w_m fixed at 0.4418. Quantitative comparisons of these visual feature extractors are presented in section 4.1.

3.2. Object Clustering

At this point, we generate 3D masks for each object detected in an image leveraging depth information. To determine whether to initialize a new object or merge existing 3D masks, we compute their spatial and visual similarities.

Given a 3D mask i extracted from image I_t, represented by point cloud $P_{t,i}$ and feature vector $f_{t,i}$, we compare it against an existing object j described by point cloud P_j and feature vector f_j. We use two similarity measures:

Visual Similarity: Cosine similarity between feature vectors

$$\text{Visual Similarity} = \frac{f_{t,i} \cdot f_j}{|f_{t,i}| \cdot |f_j|}/2 + 1/2 \tag{4}$$

Spatial Similarity: Fraction of points in $P_{t,i}$ that have a close neighbor in P_j

$$\text{Spatial Similarity} = \frac{\left|\left\{q \in P_{t,i} : \min_{p \in P_j} \text{dist}(q,p) \leq \delta\right\}\right|}{|P_{t,i}|} \tag{5}$$

Merging criteria: If the sum of the similarities exceed a threshold and represents the highest similarity, we merge the masks. Otherwise, we create a new object. The merged object's semantic vector is computed as a moving average of visual features, enabling the construction of a semantically rich 3D map similar to ConceptGraphs [2].

3.3. 3D Scene Graph Generation

We segment rooms in the entire object-centric point cloud of the scene by taking its Bird's Eye View (BEV) projection. Unlike previous approaches like HOV-SG [12], we encountered challenges in segmenting room regions using watershed algorithms due to the inherent open spatial characteristics of rooms in the HM3D [10] dataset. So, we manually identified the room boundaries in our 3D scene representation. Object centroids are then used to systematically assign objects to their respective segmented rooms.

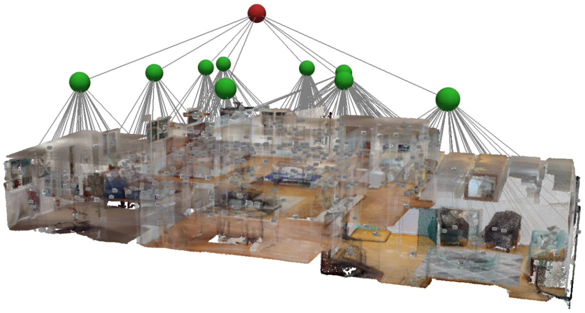


Figure 2. Scene graph generated for a house space in the HM3D [10] dataset (id: 00824-Dd4bFSTQ8gi). Green nodes represent rooms, each connected to the corresponding objects present in that room via the grey edges. The red node represents the overall building.

We then construct a graph with three types of nodes: rooms, objects, and 3D points. The graph's edges encode the hierarchical "contained in" relationships, specifically connecting 3D points to their parent objects and objects to the rooms they are contained in. Figure. 2 illustrates our scene graph in the HM3D [10] dataset.

3.4. Audio Query

The features of each object are aligned in the CLIP feature space. Consequently, the CLIP text and image encoders can be used for searching objects via text or image queries in downstream tasks. Additional modalities could be used for queries, if they can be embedded in the CLIP feature space. ImageBind [1] proposes an approach to learn a joint embedding that co-exists with the OpenCLIP [9] feature space, spanning six different modalities: images, text, audio, depth, thermal, and IMU data. We utilized the audio encoder from ImageBind [1] to encode audio queries into the CLIP feature space. Since the joint embedding space of ImageBind [1] already co-exists with OpenCLIP [9] feature space, this lets us use the object features calculated from OpenCLIP [9] without modifications.

3.5. Inference Details

One of the major highlights of the OpenScene [8] work is the use of arbitrary text queries for scene understanding. We enable the same with our scene graph. Our inference pipeline uses a Large Language Model (LLM) to transform natural language queries into structured scene details. Specifically, we use the Qwen2.5-72B-Instruct model via the HuggingChat API to parse queries and extract semantically meaningful room and object details in ['room', 'object'] format. For example, a query like "A lamp in the living room" is parsed to ["living room", "lamp"], precisely extracting the specific object and location. More open-ended queries such as "all chairs" result in [None, "chairs"], triggering a comprehensive search across all rooms. We match the CLIP embeddings of these parsed query components against the scene graph's node features to first identify the room level, and then the object in that particular room.

4. Evaluations

4.1. 3D Semantic Segmentation

To assess the zero-shot semantic understanding capabilities, we evaluate the performance on the Replica [11] dataset. This allows us to compare the feature extraction methods of ConceptGraphs [2], ConceptFusion [4], and HOV-SG [12].

Given a predefined set of class names, the semantic label for each point is determined by computing the similarity between the fused semantic feature of its associated object node and the CLIP text embeddings of the phrase "an image of {class}." following the approach in ConceptGraphs [2]. Each point is assigned the label of the class with the highest similarity. For evaluation, we perform a bidirectional nearest-neighbor point association between predicted and ground truth point clouds (GT → Prediction and Prediction → GT). We report two primary semantic segmentation metrics, class mean recall (mAcc) and frequency weighted mean intersection over union (F-mIoU) in Table 1

From Table 1, we observe that HOV-SG [12] achieves comparable mAcc to ConceptGraphs [2] while outperforming both methods in F-mIoU, indicating better overall segmentation quality. The performance improvement of HOV-SG [12] can be attributed to the usage of local features, which are a weighted sum of the cropped image and masked cropped image. The masked cropped image reduces background interference, and hence minimizes the influence of other objects on the target object. The lower performance of the ConceptFusion [4] method might be due to the differences in the overall pipeline. The authors of ConceptFusion [4] used SAM [5] without bounding boxes detected by a foundational model like GroundingDINO [6]. Since we used foundational models for object detection and used those bounding boxes as input to SAM [5], we obtained a 2D mask representing the whole object. Fusing the global feature with this might have caused an interference from the background objects onto the 2D masks, affecting the feature space of the 2D masks.

4.2. Object retrieval via text with scene graph and without scene graphs

3D scene graphs help represent the hierarchical topological information of a 3D map. To evaluate the effectiveness, we conducted an experiment based on object retrieval success rate as shown in Table 2. The Top-k metric indicates whether the expected object was among the first k matches predicted by the method. A search is counted as successful if at least 20% of the predicted object's points have class labels that match the expected class label.

Method	Scene ID	mAcc [%]	F-mIoU [%]
HOV-SG	room0	36.21	37.39
	room1	**47.51**	**44.11**
	room2	27.19	28.63
	office0	**33.81**	**33.68**
	office1	33.06	**21.55**
	office2	**36.81**	**56.45**
	Overall	38.24	**36.13**
ConceptGraphs	room0	**41.49**	41.87
	room1	40.47	37.59
	room2	**29.30**	**36.40**
	office0	32.51	28.92
	office1	**33.19**	18.29
	office2	33.25	47.01
	Overall	**38.66**	34.54
ConceptFusion	room0	39.65	**42.02**
	room1	40.21	36.75
	room2	25.82	36.17
	office0	32.60	28.95
	office1	29.00	18.29
	office2	32.49	47.09
	Overall	36.46	34.35

Table 1. Open-vocabulary semantic segmentation experiments on the Replica [11] dataset.

Due to compute and memory constraints, we conducted this experiment on a subset of the Replica [11] dataset instead of the HM3D [10] dataset. We selected room 0, room 1, room 2, and office 2 and used them as different rooms of a house since they had distinct semantic characteristics. After extracting ground truth labels for each room, we removed non-relevant objects such as walls, floors and ceilings. This process yielded a total of 81 object instances. We used GPT-4o to generate queries in the form ["A/an {object_name} in {room_name}", [{room_name, object_name}]. The first index of the output from GPT-4o served as the query for evaluation, and the second index provided the expected room name and expected object name. A total of 81 such queries were generated for the analysis.

Metric	Top-1	Top-2	Top-3
With Scene Graph	0.6049	0.6296	0.6914
Without Scene Graph	0.3086	0.4444	0.5556

Table 2. Comparison of query success rates with and without the use of Scene Graph structure. The Top-k metric indicates whether the expected object was among the first k matches predicted by the method.

As shown in Table 2, the scene graph representation outperforms the no-scene-graph representation across all three metrics. The reason is that the 3D scene graph representation allows us to find the best-matched room first and then search for all objects inside that room. The chance of finding the correct object is high. Note that the open vocabulary aspect of our method is restricting the maximum success rate to 0.6914, as some classes were undetected by the foundational models and thus were not considered for the CLIP feature calculation. In the scenarios where no scene graphs are used, we search through all objects in all rooms, and there could be chances we might obtain another similar object in a different room which might have a higher cosine similarity with the query.

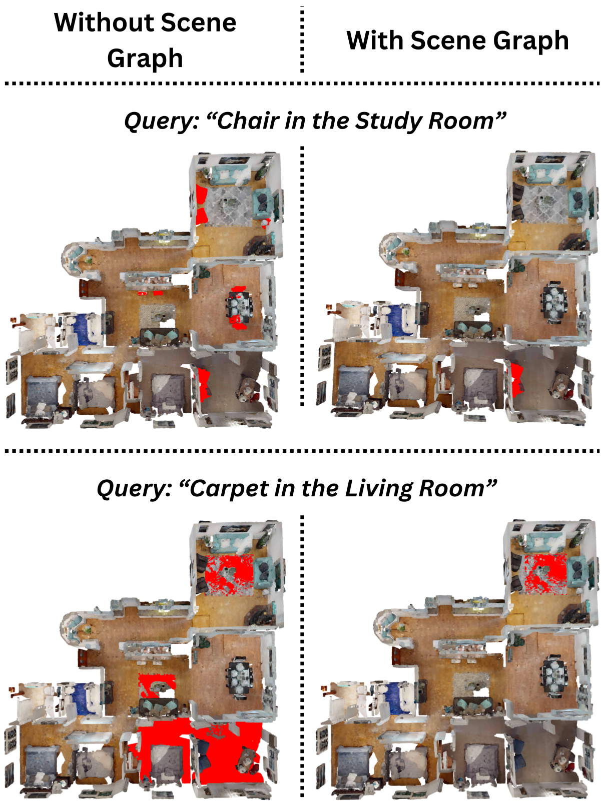

Without Scene Graph
With Scene Graph
Query: "Chair in the Study Room"
Query: "Carpet in the Living Room"

Figure 3. Two examples from the HM3D [10] dataset comparing the method without and with scene graph.

4.3. Qualitative Analysis

We present qualitative results to visually demonstrate the method's capabilities:

1. Scene Graph: Figure 3 shows two examples from the HM3D dataset (scene id: 00824-Dd4bFSTQ8gi) comparing the method without scene graph and with scene graph. Using Scene graphs helps localize the objects in the right location since it provides a higher-level topological understanding.
2. We show the method's inference results across different input modalities - text, image and audio:
   (a) Text Queries: Fig. 4a shows the method's semantic understanding capabilities through diverse text queries representing affordances, materials, objects, and rooms.
   (b) Image Queries: Fig. 4b shows how we could use images to perform object retrieval. For instance, using images like an orange towel helps identify the towel in the washroom. Additional examples involving the images of a bathtub and plants showcase its performance for image based inference.
   (c) Audio Queries: Fig. 4c depicts the scene understanding when we use an audio clipping of the sound of the toilet being flushed, snoring and an alarm clock. It neatly identifies the corresponding objects like toilet and beds.

5. Summary

We have shown a simple pipeline for building hierarchical, open-set, multi-modal map representations, drawing inspiration from recent cool works. These representations show great potential in scene understanding for robots. A promising future direction is extending this approach to SLAM systems with back end optimization, enabling real-time mapping systems that are open-set and hierarchical.

However we also acknowledge that further work needs to be done in our work for 2 main areas: robust room segmentation and quantitative evaluations for audio and image based queries.

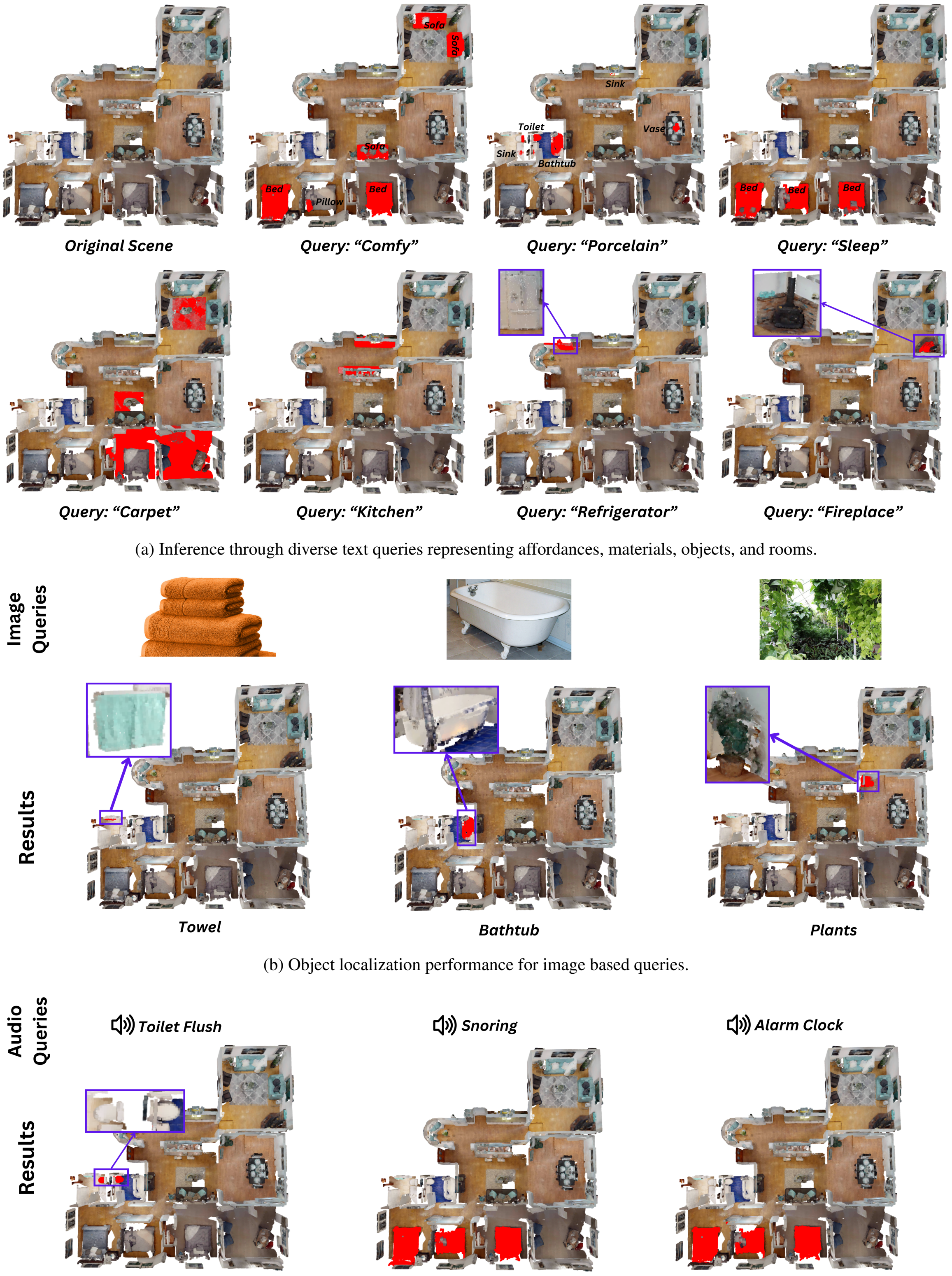
References

[1] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023. 1, 3

[2] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028, 2024. 2, 3, 4

[3] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980, 2022. 1

[4] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B.



(a) Inference through diverse text queries representing affordances, materials, objects, and rooms.

(b) Object localization performance for image based queries.

(c) Inference for audio based queries.

Figure 4. Qualitative performance of the method across multiple modalities - text, image and audio on the HM3D [10] dataset (scene id: 00824-Dd4bFSTQ8gi)

Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems (RSS)*, 2023. 2, 4

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 1, 2, 4

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 4

[7] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1

[8] Songyou Peng, Kyle Genova, Chiyu “Max” Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–824, 2023. 2, 3

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[10] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 3, 4, 5, 6

[11] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 4

[12] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024. 2, 3, 4

[13] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1724–1732, 2024. 2