RoboAuditor: Goal-Oriented Robotic System for Assessing Energy-intensive Indoor Appliance via Visual Language Models

Weijia Cai
The University of British Columbia
Vancouver, BC, Canada
andycai@student.ubc.ca

Lei Huang
The University of British Columbia
Vancouver, BC, Canada
lei.huang@ubc.ca

Zhengbo Zou
The University of British Columbia
Vancouver, BC, Canada
zhengbo@civil.ubc.ca

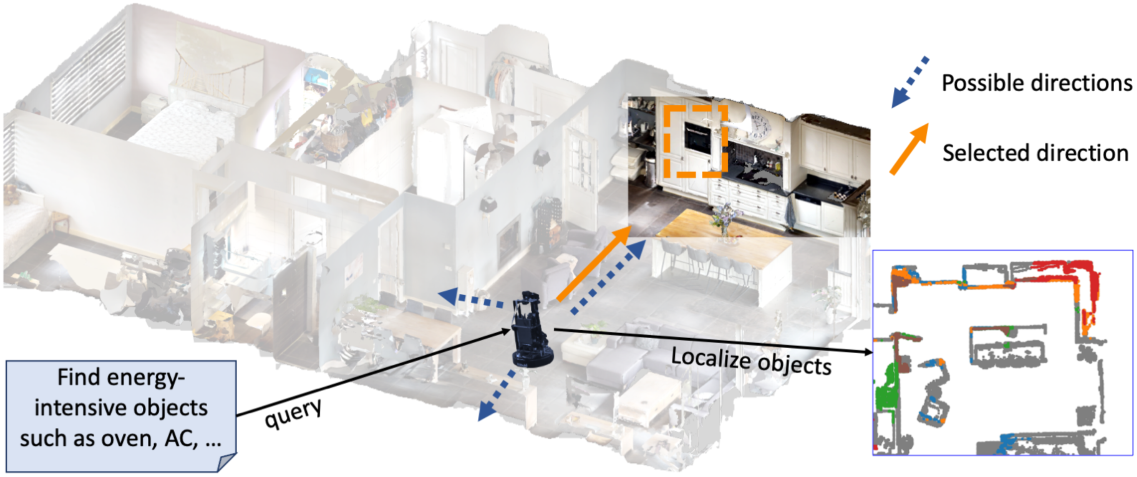

Figure 1: Overview of RoboAuditor. RoboAuditor receives queries of a list of energy-intensive objects, plans navigation towards the queried objects, and then localizes detected objects in 2D semantic map.

ABSTRACT

Energy auditing is a crucial step in building retrofitting to enhance building energy efficiency. However, auditing tasks, such as profiling energy-consuming appliances in buildings, rely heavily on human inspectors, resulting in a time- and capital-intensive process. To this end, we propose an autonomous robotic system, dubbed RoboAuditor, for identifying and localizing energy-intensive appliances in buildings given text queries from humans. RoboAuditor utilizes visual language models to predict relevance scores between text queries and observed images for goal selection in robot navigation. It then automatically identifies and localizes queried appliances while self-navigating with efficient navigational strategies. For evaluation, we deploy the proposed robotic system on a wheeled robot equipped with an RGB-D camera and run auditing tests in 12 residential buildings in 3D simulation. These buildings exhibit diverse room counts, appliance quantities, and navigable areas, and they all feature energy-intensive appliances, such as air conditioners, heaters, dishwashers, and refrigerators. We conduct two groups of experiments: the first group uses the relevance score, and the second serves as a control group without the relevance score. Results demonstrate that RoboAuditor detects queried appliances and accurately localizes their positions in buildings with an average success rate of 68.05%, showing a significant margin of 6.8% higher than the control group.

CCS CONCEPTS

• **Applied computing** → Engineering; • **Computing methodologies** → Artificial intelligence; • **Computer systems organization** → **Robotics**.

KEYWORDS

Energy auditing, robotics, deep learning, visual-language model

ACM Reference Format:
Weijia Cai, Lei Huang, and Zhengbo Zou. 2023. RoboAuditor: Goal-Oriented Robotic System for Assessing Energy-intensive Indoor Appliance via Visual Language Models. In *The 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '23), November 15–16, 2023, Istanbul, Turkey*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3600100.3623739

1 INTRODUCTION

Buildings account for around 70% of the electricity consumption in the U.S. [21]. According to the U.S. EIA [14], energy consumption

BuildSys '23, November 15–16, 2023, Istanbul, Turkey

ACM ISBN 979-8-4007-0230-3/23/11…$15.00
https://doi.org/10.1145/3600100.3623739

and greenhouse gas (GHG) emissions of the building sector are projected to grow through 2050. To curb this growth, government agencies and professional entities roll out incentives for building owners and operators to improve building energy efficiency. For example, retrofitting aged buildings improves their energy performance, since 85% of residential buildings in the U.S. were built before 2000, predating modern energy codes [12]. Many of these buildings were designed in a less efficient manner and constructed using degraded materials [61], causing excessive energy consumption compared to new buildings [35, 41]. Furthermore, to reduce peak electricity demand and avoid large scale blackouts, energy providers in large cities incentivize building owners to save energy through demand response (DR) programs [1, 9], which require buildings to reduce their energy consumption to a preset target level during peak hours. For another example, building energy performance certification programs, such as ENERGY STAR [5], are established to encourage developers to build energy-efficient and sustainable buildings [3], keeping building sustainability at the forefront from the outset.

Among these endeavors, energy auditing is usually a critical initial step for profiling accurate building details. Energy audits (i.e., energy survey) comprehensively assess how the energy consumption is distributed across facilities during building operation and suggest potential areas for improvements [58]. The current practice of building energy audits involves recruiting auditors to assess buildings in person, making the overall process time- and labor-intensive [4]. For example, a common task of energy audits is to create a datasheet of appliances, such as air conditioners, dishwashers, refrigerators, and lighting systems [46]. However, the significant capital and time cost could impose negative incentives on property holders regarding audits and retrofits [18]. As such, automated audits are of paramount importance for reducing the capital and time cost [30, 44]. Recent advancement in building information modeling (BIM) enables fast queries of building objects and appliances. For example, Kim and Peavy [28] used BIM and the Industry Foundation Classes (IFC) to create a semantic building world, from which robots could query information for operations. Although pervasive in newly-constructed buildings, BIM still remains an elusive tool for aged buildings, and the quality of models is often brought into question during audits, leading to rework of the auditors. On the other hand, researchers have utilized robots to assist with energy auditing and retrofitting. For example, Mantha et al., [43] developed a framework for retrofit decision making with data collected by an autonomous mobile robot. But their work focused on collecting ambient parameters, such as temperature and humidity, for building energy modeling.

To this end, we aim to develop an automated in-situ robotic auditing system, dubbed RoboAuditor, that can deliver certain tasks in energy audits. Specifically, RoboAuditor includes four modules, namely, a Simultaneous Localization and Mapping (SLAM) module for environment reconstruction and robot localization, a frontier detection module for detecting navigational goals, a relevance score mapping module that utilizes visual-language models for detecting queried objects in the observed area and scoring the navigational goals, and a goal assigner module for determining subsequent navigational objectives. In developing RoboAuditor, the main contributions of this study are highlighted as follows:

- We develop a robotic solution for the identification and localization of energy-intensive appliances in buildings.
- We utilize zero-shot visual language models for robots to understand the text queries from humans without additional training, and perform efficient robot navigation.
- We conduct experiments in 12 residential buildings at three scales in realistic virtual environments captured using high-resolution 3D scans.
- We achieve efficient path planning and accurate localization of the queried energy-intensive objects.

2 BACKGROUND

2.1 Goal-Oriented Exploration

To facilitate robots' capability to autonomously explore uncharted environments, SLAM algorithms serve the dual function of reconstructing a map of the unknown environment and locating positions of the robot within the reconstructed map [42]. One of the main challenges is that the robot only receives partial information about the environment while having to produce well-explored 2D or 3D maps [64]. Current exploration strategies vary based on different data structures of goals (e.g., frontier points and image) [7, 64, 68] and criteria of the exploration goal selection (e.g., energy used by the robot, distance traveled, and information gain) [59, 60, 62, 66].

For frontier-based exploration methods, there are three basic types of space: open space, occupied space, and unknown space in a 2D occupancy map [64]. Subsequently, frontiers are described as a collection of unknown points that possess at least one neighboring point classified as open space. Frontier-based exploration strategies utilize the detected frontiers as navigational goals to conduct path-planning tasks. To improve the exploration efficiency of frontier-based methods, researchers proposed rapidly-exploring random tree (RRT) methods [26, 27, 34], which only visit a collection of connected frontiers in a random tree expanded in real-time. Recent studies [6, 69] improve RRT methods by constructing topological graphs to reduce resampling and expedite exploration. However, these methods require denser sensor input such as LiDAR to ensure robustness [63].

Recent works propose a novel visual navigation task named object goal navigation that requires a robot to explore the environment given an image target as a goal to reach [7]. Most works for object goal navigation use learning-based exploration methods such as Deep Reinforcement Learning (DRL) [7, 17, 20]. For example, Chaplot et al. [7] proposed a goal-oriented DRL-based framework that takes real-time semantic maps and visual observations (i.e., RGB-D images) as input and predicts navigational goals on a 2D map. Georgios et al. [20] adopted a similar pipeline but used a pre-trained semantic map predictor. These DRL-based methods learn the alignment between goal objects and path-planning strategies. However, they suffer from the drawback that a significant number of training steps is needed to obtain an acceptable performance, which increases the computational cost [22, 45].

2.2 Language-image Pre-training for Robot Navigation

Pre-training methods that learn directly from unstructured texts have been prevailing in Nature Language Processing (NLP) [11, 13, 24, 49]. Models pre-trained under self-supervised schemes are proved to be highly effective in downstream tasks, such as text completion [13] and question answering [50]. Inspired by language pre-training, recent studies have demonstrated the efficacy of visual pre-trained models in enhancing their performance in common computer vision tasks [10, 15, 23].

While the effectiveness of pre-trained models has been established in tasks related to both vision and language individually, recent attention has increasingly shifted towards the exploration of representation learning that encompasses multiple modalities (e.g., image and text) [37, 48, 53, 67]. Specifically, visual language models (VLM) aim to process and link information from image, video, and text. For example, Radford et al. [48] proposed a cross-modality pre-training approach (CLIP) that learns visual representation from language supervision. CLIP is trained on crowd-sourced internet-scale data whose quality remains uncontrollable. Li et al. [37] attempted to close this gap by developing a dataset bootstrapping method (BLIP) that generates synthetic image captions and automatically filters both noisy web text and the synthetic captions. BLIP shows state-of-the-art performance in image-text matching, image captioning, and other vision language understanding tasks. Furthermore, VLM has been leveraged in open-vocabulary object detection, such as DetCLIP [65] and Grounding-DINO [39]. In this paper, we utilize Grounding-DINO [39] to measure the similarity between a queried text and images observed by the robot, while localizing the queried objects in images.

The advancements in VLM create significant opportunities in the field of robot navigation [2, 25, 38, 56]. For example, Shah et al. [56] proposed a modular framework (LM-Nav) that enables mobile robots to follow instructions provided in natural language and navigate to instructed locations. LM-Nav utilizes large pre-trained models to ground visual observations from text instructions without additional training. Gradre et al. [19] accomplished object navigation based on language cues by constructing a relevance map for the object category via CLIP and GradCAM [40]. Similarly, Huang et al. [25] constructed a visual language map using a pre-trained language-guided segmentation model [36] for robot navigation searching for multiple objects. These prior works are encouraging, since they show pre-trained VLMs are effective in producing queryable scene representation for object goal navigation. However, most of them do not explicitly use the extracted information from VLMs in robot navigational strategies. In this paper, we focus more on utilizing the features produced by pre-trained VLMs for frontier selection in real-time goal-oriented exploration.

3 METHODOLOGY

In this section, we develop RoboAuditor, a modular system consisting of an RGB-D SLAM module, a frontier detector, a relevance score mapping module, and a goal assigner [33, 59]. We develop this system under Robot Operation System (ROS) [47], which enables efficient communication among each module. Figure 2 shows an overview of our approach. The RGB-D SLAM module consumes RGB-D frames and produces real-time camera poses and a 2D occupancy grid. A frontier detector takes in the 2D occupancy grid and generates a series of frontier points. The relevance score mapping module then predicts the relevance score between the current image observation and the given text query. Then a relevance score mapper projects the score for the observed area on a top-down map. Finally, the goal assigner selects a frontier point with the highest revenue value, regarding relevance score, information gain, and distance cost, as the navigational goal.

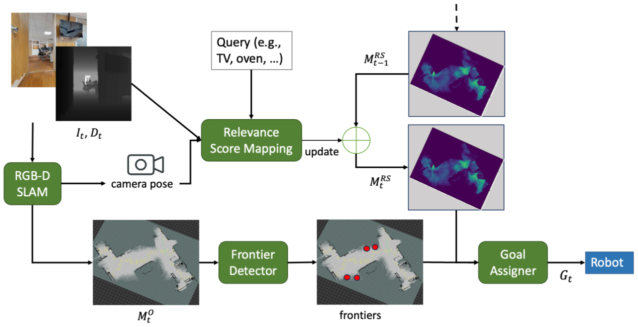

Figure 2: Overview of RoboAuditor's modules. At time step t, RGB-D SLAM module takes an RBG image I_t and a depth image D_t as inputs and estimates a 7 DoF camera pose $Pose_t$ and a 2D occupancy grid $\mathcal{M}_t^O$. A frontier detector computes feasible frontiers from $\mathcal{M}_t^O$. A relevance score mapping module takes camera pose, I_t, D_t, and the relevance score map from the previous time step $\mathcal{M}_{t-1}^{RS}$ as inputs to update the current relevance score map $\mathcal{M}_t^{RS}$ at time step t. Subsequently, a goal assigner continuously utilizes $\mathcal{M}_t^{RS}$ to rank the currently detected frontiers and send the robot the best frontier point G_t to navigate.

3.1 RGB-D SLAM and Frontier Detection

The RGB-D SLAM module is used to track the robot's pose (e.g., position and rotation) and build a 3D map from the visual observations including RGB and depth images. The 3D map is then projected to the ground as a 2D occupancy grid for frontier detection.

3.1.1 Pose Estimation and Tracking. The RGB-D SLAM module produces poses of robots for registering the detected objects in the global frame. In RGB-D SLAM, pose estimation is obtained by Visual Odometry (VO) [55]. VO estimates the egomotion (3D movement of a camera within an environment) of a robot given only visual observations (e.g., RGB and depth images). Essentially, VO estimates the transformation matrix $T_{t-1,t}$ between two consecutive poses, X_{t-1} and X_t, of a robot corresponding to their input RGB frames I_{t-1} and I_t, as well as depth frames D_{t-1} and D_t. This transformation between X_{t-1} and X_t can be expressed as $X_t = T_{t-1,t}X_{t-1}$.

3.1.2 Mapping. The RGB-D SLAM module also produces 2D/3D mapping of unknown environments, which is used as the input for frontier-based exploration. Mapping consists of two steps: 1) local map creation and 2) global map assembling [32]. As a depth sensor is available in RGB-D SLAM, a local map with respect to the camera coordinate system can be obtained by projecting the depth information to a local point cloud $m_t^{'}$. The computation is expressed by equation 1.

$$m'_t = \psi(u,v) = \begin{cases} z = d \times \lambda \\ x = (u - c_x) * z/f_x \\ y = (v - c_y) * z/f_y, \end{cases} \tag{1}$$

where ψ is the function for transforming depth image to a point cloud, (u, v) denote the position in terms of pixel coordinates of the image plane, f_x and f_y are focal lengths for x and y axis, c_x and c_y are principal point offsets, and $\lambda = 1.0$ is the depth scale. The camera parameters including f_x, f_y, c_x, and c_y are assumed to be known.

Global map assembling is computed and optimized along with VO. It registers local maps with valid estimated odometry into the global frame. As mentioned in 3.1.1, VO estimates the transformation matrix $T_{t-1,t}$ between two consecutive frames. $T_{t-1,t}$ is then used to register the local map m'_t by equation 2.

$$m_t = T_t m'_t = (\prod_{i=1:t} T_{t-1,t}) m'_t, \tag{2}$$

where m_t is the registered point cloud in the global frame, T_t is the estimated robot pose at time step t, which can be computed by multiplying a series of transformation matrices. Note that m'_t is transformed to homographic form. As the odometry is optimized locally on two consecutive frames, estimation errors will accumulate over time. To mitigate this issue, loop closure and graph optimization techniques [31] are applied to optimize the pose estimation and map registration over all time steps. The constructed 3D global point cloud is projected to the ground as the 2D occupancy grid for frontier detection.

3.1.3 Frontier Detector. Within the mapping process of the RGB-D SLAM module, frontiers are detected using a real-time 2D occupancy grid derived from projecting the registered 3D map onto the ground. We implemented a naive frontier detector (NFD). As defined in 2.1, a frontier point is an unknown point with at least one neighbor belonging to open space. Based on the definition, NFD samples points from borders between open space and unknown space, shown in Figure 3.

As we can see in Figure 3, collision issues may occur when the sampled frontier points are too close to obstacles. Therefore, we apply a filter node [33, 59] to select valid frontier points. A frontier point is valid if 1) its neighboring area (i.e., a fixed box area centered at the frontier point) does not contain obstacles, and 2) its information gain (i.e., the number of unknown points in the neighboring area) is greater than a predefined threshold.

3.2 Relevance Score Mapping

In this section, we utilize pre-trained VLMs to compute relevance scores between queried objects and observed images. A relevance score map (RSMap) is used to indicate which part of the environment has a higher probability of containing the queried object. RSMap is created using the predicted relevance score, depth images, and the estimated poses from RGB-D SLAM.

3.2.1 Relevance Score Predictor. Pre-trained VLMs align images and texts with similar semantics. Figure 4 shows a general pipeline of Relevance Score Predictor (RSP). Image input I_t contains the queried object (i.e., TV). A pre-trained image encoder extracts a feature vector φ_t^I with respect to the image I_t. A text encoder extracts features with respect to the queried objects (e.g., "oven", "AC", and "TV"). Then the Grounding-DINO module takes both text features (e.g., φ^{oven}, φ^{AC}, and φ^{TV}) and image features φ_t^I as inputs and proposes locations of the detected objects. Each object's location has a relevance score corresponding to every object category. Since the image encoder and the text encoder are pre-trained for text-image alignment, the text feature is more relevant to the image feature with the same semantics. For each time step, RSP only outputs the detection with the highest score.

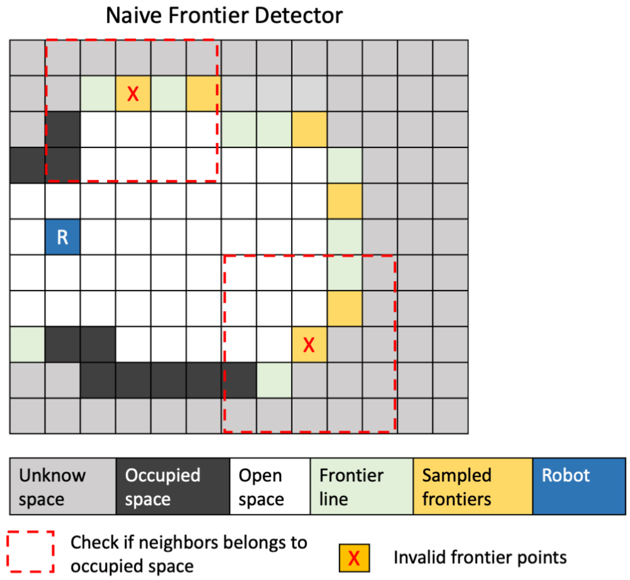

Figure 3: NFD samples frontier points from the frontier line. Each frontier point will be checked if there is a neighbor belonging to occupied space. The sampled frontiers with "X" stands for invalid frontier points due to possible collisions.

3.2.2 Relevance Score Mapper. Once the relevance score is obtained from VLM, a relevance score mapper assigns the score to a local area in the top-to-down RSMap. The pipeline of this process is shown in Figure 5. First, a depth image D_t is projected to a local point cloud using equation 1. Then the predicted bounding box b_t is used to crop the detected object in D_t. Later, the cropped local point cloud is registered in the global frame with the camera pose from RGB-D SLAM module using equation 2. This patch of point cloud $P_t = (x_t, y_t, z_t)$ is then projected to the ground and transformed to pixel coordinates (u_t, v_t) on the 2D RSMap by equation 3.

$$P_t^{RS} = (u_t, v_t) = \begin{cases} u_t = \lfloor (x_t - x_c)/s \rfloor + 1 \\ v_t = \lfloor (y_t - y_c)/s \rfloor + 1, \end{cases} \tag{3}$$

where (x_c, y_c) is the map origin in the global frame of the RSMap; $s = 0.03m$ is the grid size of RSMap, which defines its pixel size; $\lfloor \cdot \rfloor$ is the floor operator. $\mathcal{M}_t^{RS}$ is the relevance score map that is updated by merging the former RSMap $\mathcal{M}_{t-1}^{RS}$ and the current update P_t^{RS}. It is worthwhile to note that $\mathcal{M}_t^{RS}$ is a probability distribution of

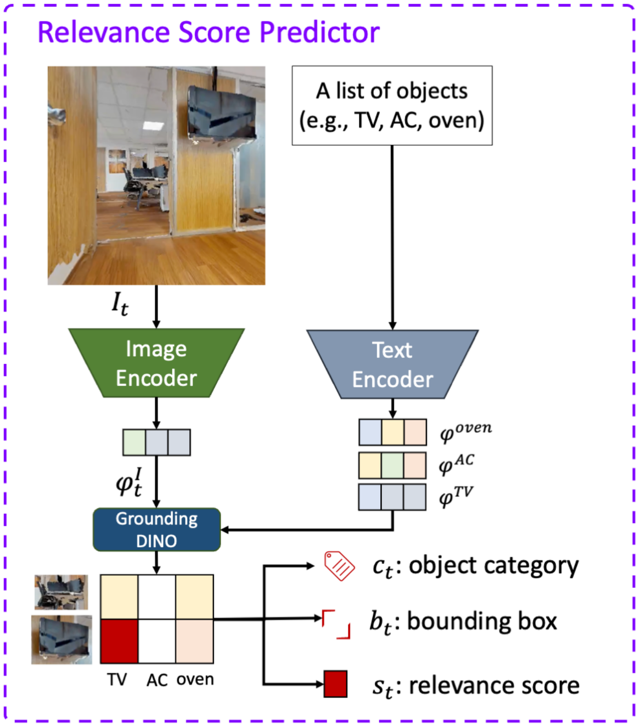

Figure 4: Relevance score predictor. I_t is an observed image at time step t, s_t is its predicted relevance score, b_t is its predicted bounding box for the detected object with the highest score s_t, c_t is the predicted object category.

the location of the queried objects. This property will be used to help robot navigation and object localization.

3.3 Goal Assigner

We implement a goal assigner that ranks the detected frontier points and determines the current navigational goal based on Algorithm 1. First, the goal assigner will wait until it receives $\mathcal{M}_t^O$ from the RGB-D SLAM module and $\mathcal{M}_t^{RS}$ from the relevance score mapping module. Next, an initial frontier set is extracted from $\mathcal{M}_t^O$. Then, the goal assigner ranks the frontier points by assigning each of them a revenue value, which is computed using equation 4.

$$\mathcal{R}(x_{fp}, x_r) = \alpha \mathcal{I}(x_{fp}) \times \mathcal{M}_t^{RS}(x_{fp}) - C(x_{fp}, x_r), \quad (4)$$

where x_{fp} is a frontier point; x_r is the current pixel position of the robot; $\alpha = 10$ is a scaling factor; $\mathcal{I}$ is the information gain of x_{fp}, it computes the area of unknown space centered at x_{fp}; C is a cost function that outputs a normalized distance between x_{fp} and x_r.

The relevance score is used to weigh the importance of a frontier point. After renewing the revenue values of all frontier points, the goal assigner will select the frontier point with the highest revenue value and send it to the robot as the current navigational goal. Once the robot received the navigational goal, a local path planner is deployed to plan a collision-free path from the current robot position to the goal position. In this paper, we use Dynamic Window Approach (DWA) planner [16] which requires fewer computational resources. These processes continue until there is no valid frontier point in the occupancy grid.

Algorithm 1 Goal Assigner

```
Definition:
    $x_r = (u_r, v_r)$: robot's pixel position in global frame.
    $\mathbb{F}$: frontier set.
    $\mathcal{M}_t^O$: 2D occupancy grid received at time $t$.
    $\mathcal{M}_t^{RS}$: 2D relevance score map received at time $t$.
    $\mathcal{D}(\cdot)$: frontier detector.
    $\mathcal{R}(\cdot)$: revenue function.
    $C(\cdot)$: cost function.
    $\mathbb{R}$: revenue record.
Output:
    $G_t$: goal positions.

while $SizeOf(\mathcal{M}_t^O) == 0$ do
    try to receive initial $\mathcal{M}_t^O$ from 3.1
end while

$\mathbb{F} \leftarrow \mathcal{D}(\mathcal{M}_t^O)$.

while $SizeOf(\mathbb{F}) > 0$ do
    update $\mathcal{M}_t^O$ from 3.1.
    update $\mathcal{M}_t^{RS}$ from 3.2.
    update $\mathbb{F} \leftarrow \mathcal{D}(\mathcal{M}_t^O)$
    $\mathbb{R} = \varnothing$
    for $f \in \mathbb{F}$ do
        update $x_r$ by equation 3.
        obtain $f$'s revenue value $r = \mathcal{R}(f, x_r)$ by equation 4.
        $\mathbb{R} \leftarrow \mathbb{R} + \{r\}$
    end for
    $winner\_id = \underset{id}{\operatorname{argmax}}(\mathbb{R})$
    $G_t \leftarrow \mathbb{F}[winner\_id]$
    send $G_t$ to robot.
end while
```

4 EVALUATION AND DISCUSSION

The goal of our approach is to 1) perform efficient goal-oriented exploration strategies for indoor building environments, and 2) identify and localize multiple energy-intensive objects in indoor environments utilizing the zero-shot Grounding-DINO model. To evaluate these capacities, we conduct a series of experiments to assess multiple objects in 12 residential buildings at different scales in simulation.

4.1 Experiment Setup

4.1.1 Simulator and Dataset. We evaluate RoboAuditor in the Habitat platform [54, 57], which contains a large number of 3D indoor environments collected from real-world buildings. In this paper, we use HM3D with semantic annotations [51, 52] in which we further select 12 residential buildings at three different scales (i.e., small, medium, and large). The detailed information on the test buildings is shown in Table 1. "hs1" stands for "house small 1", "hm1" stands

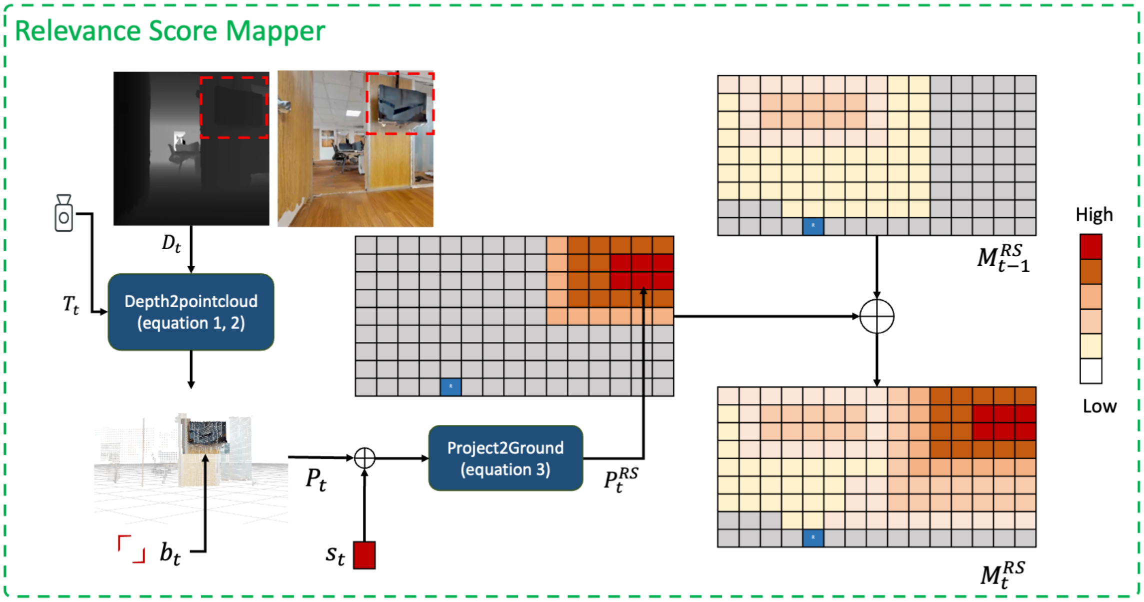

Figure 5: Relevance score mapper. D_t is observed depth image at time step t, T_t is the corresponding pose estimated by RGB-D module, P_t is the cropped point cloud in the global frame. P_t color-coded by s_t is projected to ground as P_t^{RS}.

Table 1: Test Dataset Description

ID	Scale	NumOfRoom	NavigableArea	ObjectType					
			(m^2)	TV	oven	Refrigerator	dishwasher	AC	heater
hs1	small	5	58.17	1	1	1	1	0	0
hs2	small	4	56.34	2	1	1	1	0	0
hs3	small	5	32.93	1	1	1	1	0	0
hs4	small	5	36.09	0	1	0	0	1	0
hm1	medium	8	78.73	1	1	0	1	0	0
hm2	medium	6	51	1	1	1	0	0	3
hm3	medium	7	176.62	2	1	1	0	0	0
hm4	medium	7	65.57	1	1	0	0	1	0
hl1	large	10	57.53	1	1	0	1	1	0
hl2	large	13	161.42	3	2	1	1	0	0
hl3	large	17	184.47	4	1	1	0	0	0
hl4	large	17	260.86	1	1	1	0	0	0

for "house medium 1", and "hl1" stands for "house large 1". The selected buildings are divided into three groups according to the number of rooms and navigable areas. These buildings are also required to contain different types of energy-intensive objects (i.e., air conditioner, heater, TV, dishwasher, oven, and, refrigerator). A wheeled robot (LoCoBot) with an RGB-D camera is used as the embodiment in the experiments. To integrate Habitat platform with ROS framework, we utilize ROS-X-Habitat [8] to provide a bridging service.

4.1.2 Evaluation Metrics. We use three metrics to quantitatively evaluate the performance of RoboAuditor: 1) average Success Rate (aSR) that encompasses all objects, shown in equation 5, 2) traveled trajectory length *Len.*, and 3) coverage (*Cov.*).

$$aSR = \frac{1}{N}\sum_{k=1}^{N} iSR(k) = \frac{1}{N}\sum_{k=1}^{N}\frac{N_s^{(k)}}{N_{all}^{(k)}}, \tag{5}$$

where $iSR(k)$ is individual success rate for object category k, $N_s^{(k)}$ is the number of detected instances of object category k, $N_{all}^{(k)}$ is the total number of instances of object k, and N is the number of object categories.

We use *Len.* and *Cov.* to test the navigational efficiency of the robot. *Len.* is defined as the total traveled distance and *Cov.* is defined as the percentage of the navigated area over the total navigable area.

4.1.3 Experiments. We conduct two groups of experiments: group I uses RSMap and NFD, group II uses constant score (i.e., score equals 1) and NFD as the control group, indicating that group II does not have human text query information for navigation. Group I and group II are used to test the effectiveness of RSMap in robot navigation and object localization. For group I, five sets of experiments are carried out for all 12 buildings. The evaluation results are derived from the average values obtained across these five sets of experiments.

4.2 Results and Discussions

4.2.1 Quantitative Evaluation. We quantitatively evaluate RoboAuditor by three metrics: aSR, *Len.*, and *Cov.*. The results of group I with RSMap are shown in Table 2. RoboAuditor achieves an average success rate of 68.05% over all scales of buildings, which achieves state-of-the-art performance [25, 40]. The average success rates exhibit variations based on the scale of the buildings. Small-scale buildings demonstrate the highest average success rate of 69.18%, while large-scale buildings show the lowest success rate of 65.99%. It is expected because localizing objects naturally gets more challenging when the navigable area in buildings increases. For the robots' traveled distance (*Len.*), it is not surprising that its value rises as the building scale increases, ranging from an average of 44.22 m in small-scale buildings to an average of 121.82 m in large-scale buildings. *Cov.* shows a similar pattern with aSR over all scales of buildings. The average coverage for small-scale buildings is the highest, reaching 91.71%, while the average coverage for large-scale buildings is the lowest (73.74%).

Table 2: Quantitative results for Group I (w/ RSMap)

ID	Scale	aSR ↑	*Len.* ↓	*Cov.* ↑
		(%)	(m)	(%)
hs1	small	65.33	68.08	100
hs2	small	72.22	36.92	100
hs3	small	72.50	31.05	90.64
hs4	small	66.67	40.83	76.18
Average		69.18	44.22	91.71
hm1	medium	90.00	156.81	71.41
hm2	medium	66.25	157.78	90.70
hm3	medium	70.00	80.38	87.26
hm4	medium	50.00	67.68	100.00
Average		69.06	115.66	87.34
hl1	large	59.82	128.55	76.49
hl2	large	65.00	120.45	82.07
hl3	large	50.00	73.94	72.23
hl4	large	88.89	164.35	64.18
Average		65.99	121.82	73.74
Overall average		68.05	93.90	84.26

Table 3: Quantitative results for Group II (w/o RSMap)

ID	Scale	aSR ↑	*Len.* ↓	*Cov.* ↑
		(%)	(m)	(%)
hs1	small	48.21	43.13	100.00
hs2	small	83.33	36.92	100.00
hs3	small	69.17	31.05	93.55
hs4	small	79.17	40.83	100.00
Average		69.97	37.98	98.39
hm1	medium	63.89	64.09	86.43
hm2	medium	60.00	48.72	93.83
hm3	medium	65.00	40.23	79.80
hm4	medium	66.67	67.68	75.67
Average		63.89	55.18	83.93
hl1	large	60.40	80.85	82.44
hl2	large	55.55	102.04	96.03
hl3	large	54.17	41.52	87.78
hl4	large	29.17	77.82	78.36
Average		49.82	75.56	86.15
Overall average		61.23	56.24	89.49

The results of group II are shown in Table 3. When the building scale increases, group II shows a similar pattern to group I with respect to the reduction of average success rate and average coverage percentage, and the growth of average traveled distance. The average success of group II is 61.23%, which is 6.8% lower than group I. The improvements in the average success rate differ largely for the three scales of buildings. To elaborate, the average success rate of group I outperforms the control group by 16.17% for large-scale buildings, 5.18% for medium-scale buildings, while being almost on-par with a difference of 0.79% for the small scale. This indicates the effectiveness of RSMap for identifying and localizing the queried objects, especially for larger navigable areas.

We also observe that the traveled distance and coverage are approximately proportional to the scale of the buildings for both groups. However, the average *Len.* of group I is significantly higher than group II. This is because RoboAuditor tends to explore far navigational goals with higher scores even if there are other goals nearby. Specifically, as the criteria of the navigational goal selection not only depends on relevance score but also the information gain and distance cost according to the revenue function (equation 4), there may exist a goal with high relevance score but low information gain. As a consequence, the revenue value of this goal could become relatively small compared to another goal that have a lower relevance score but higher information gain. This discrepancy leads to a higher likelihood of revisiting goals with higher relevance scores, resulting in a larger traveled distance. The revisiting behavior increases the possibility of recording queried objects during exploration.

4.2.2 Qualitative Evaluation. We visualize an example result of the localization of the energy-intensive objects in Figure 6. We generate the ground truth distribution of the queried objects by projecting the annotated 3D semantic objects to the ground. As we can observe

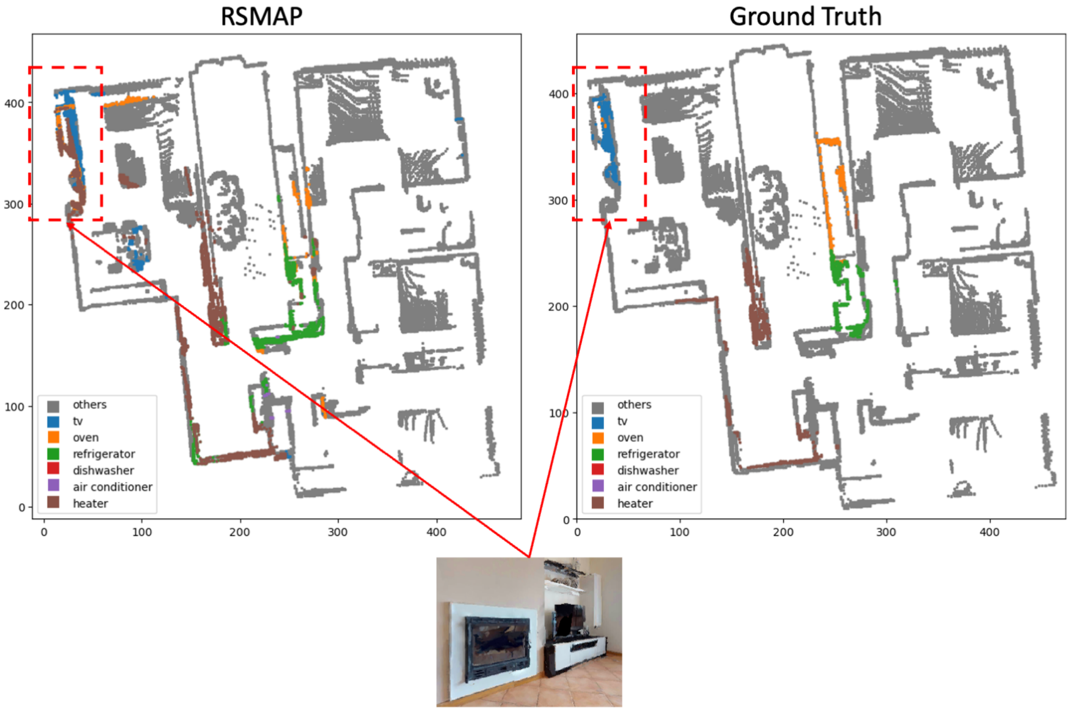

Figure 6: Qualitative results of the localization of the target energy-intensive objects. This result comes from building "hm2".

from the quantitative result, most objects (i.e., TV, refrigerator, oven, heater, and air conditioner) can be identified and localized in real time. However, there still exists false positive detection (e.g., in the left top corner). The reason for false positive detection comes from the similar appearance of objects. For instance, TV and oven are similar in their rectangular shape and color (e.g., some oven appears to be black). Besides, the text semantics between "oven" and "heater" are also similar for the pre-trained VLM. These results are as expected because we utilize the zero-shot VLM without additional training. Notably, our approach is scalable to other types of buildings, such as commercial buildings and industrial buildings, for documenting a larger repertoire of energy-intensive objects (e.g., servers, printers, scanners, and elevators) since we utilize an open-vocabulary object detector. One way to improve the object detector is to fine-tune the model with dataset under construction context.

5 CONCLUSION

In this paper, we propose RoboAuditor for efficient goal-oriented robot navigation and accurate identification and localization of energy-intensive objects in residential buildings. RoboAuditor represents the first step towards automated energy auditing using robots, which reduces time and cost while eliminating the need for prior knowledge of buildings in form of drawings or BIM. Once deployed, RoboAuditor can be used by building owners and facility managers to make informed decisions towards building improvements such as retrofitting.

RoboAuditor leverages a pre-trained Vision-and-Language Model (VLM) to align text queries with images, generating relevance scores that aid goal selection during frontier-based robot exploration. Meanwhile, RoboAuditor is capable of accurately identifying and localizing queried objects by a zero-shot open-vocabulary object detector (i.e., Grounding-DINO), and can be easily extended to include a large variety of customized queried objects and environments with no additional training.

We collected 12 residential buildings containing various energy-intensive objects at three scales from HM3D dataset. To assess RoboAuditor's performance, two groups of experiments are conducted: group I employs RSMap, while group II does not use RSMap as the control group. We obtained an average success rate of 68.05% across all buildings in group I, whereas the control group achieved an average success rate of 61.22%. Furthermore, our observations indicate that incorporating RSMap prompts the robot to revisit locations with a higher likelihood of housing energy-intensive objects, leading to an increase in the overall traveled distance.

This work has identified a few limitations that also highlight potential areas for future research. First, the localization of objects is based on bounding boxes of detection results, where all pixels within a bounding box are set to the same object category, potentially leading to false positive results within the bounding box. An alternative way is to replace Grounding-DINO with a zero-shot open-vocabulary segmentation model such as Grounded-SAM [29, 39] for more detailed segmentation of queried objects. By achieving a finer-grained identification of objects of interest, we have the potential to augment the functionality of RoboAuditor, enabling it to perform tasks such as appliance counting in future work. Second, apart from the alignment between the queried text and the observed images, the relevance score prediction can be further improved by utilizing environment context like room types [2] to ground the queried text with environment constraints (e.g., an oven is commonly located in a kitchen). Third, the RSMap for robot navigation can be extended from 2D to 3D with depth information to include more information. Moreover, RoboAuditor’s capability can be augmented by installing additional sensors. For example, thermal cameras can be plugged in to align the text queries with specific thermal leakages in buildings. Last, real-world experiments will be conducted to exam the effectiveness of RoboAuditor in our future work.

ACKNOWLEDGMENTS

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2022-04492].

REFERENCES

[1] Milad Afzalan and Farrokh Jazizadeh. 2019. Residential loads flexibility potential for demand response using energy consumption patterns and user segments. *Applied Energy* 254 (2019), 113693. https://doi.org/10.1016/j.apenergy.2019.113693

[2] Ayush Agrawal, Raghav Arora, Ahana Datta, Snehasis Banerjee, Brojeshwar Bhowmick, Krishna Murthy Jatavallabhula, Mohan Sridharan, and Madhava Krishna. 2023. CLIPGraphs: Multimodal Graph Networks to Infer Object-Room Affinities. In *ICRA2023 Workshop on Pretraining for Robotics (PT4R)*. https://openreview.net/forum?id=qm62NWMxHV

[3] Pandarasamy Arjunan, Kameshwar Poolla, and Clayton Miller. [n. d.]. EnergyStar++: Towards more accurate and explanatory building energy benchmarking. *Applied Energy* 276 ([n. d.]), 115413. https://doi.org/10.1016/j.apenergy.2020.115413

[4] Michael C Baechler. 2011. *A guide to energy audits*. Technical Report. Pacific Northwest National Lab.(PNNL), Richland, WA (United States). Retrieved 2023-06-20 from https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-20956.pdf

[5] R. Brown, C. Webber, and J.G. Koomey. [n. d.]. Status and future directions of the Energy Star program. *Energy* 27, 5 ([n. d.]), 505–520. https://doi.org/10.1016/S0360-5442(02)00004-X

[6] Chao Cao, Hongbiao Zhu, Howie Choset, and Ji Zhang. 2021. Exploring Large and Complex Environments Fast and Efficiently. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Xi’an, China, 7781–7787. https://doi.org/10.1109/ICRA48506.2021.9561916

[7] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. 2020. Object Goal Navigation using Goal-Oriented Semantic Exploration. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 4247–4258. https://proceedings.neurips.cc/paper_files/paper/2020/file/2c75cf2681788adaca63aa95ae028b22-Paper.pdf

[8] Guanxiong Chen, Haoyu Yang, and Ian M. Mitchell. 2022. ROS-X-Habitat: Bridging the ROS Ecosystem with Embodied AI. In *2022 19th Conference on Robots and Vision (CRV)*. 24–31. https://doi.org/10.1109/CRV55824.2022.00012

[9] Yongbao Chen, Peng Xu, Jiefan Gu, Ferdinand Schmidt, and Weilin Li. [n. d.]. Measures to improve energy demand flexibility in buildings for demand response (DR): A review. *Energy and Buildings* 177 ([n. d.]), 125–139. https://doi.org/10.1016/j.enbuild.2018.08.003

[10] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. 2022. When Does Contrastive Visual Representation Learning Work?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14755–14764.

[11] Andrew M Dai and Quoc V Le. 2015. Semi-supervised Sequence Learning. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/7137debd45ae4d0ab9aa953017286b20-Paper.pdf

[12] M Deru, K Field, D Studer, K Benne, B Griffith, P Torcellini, B Liu, M Halverson, D Winiarski, M Rosenberg, M Yazdanian, J Huang, and D Crawley. 2011. U.S. Department of Energy Commercial Reference Building Models of the National Building Stock. https://doi.org/10.2172/1009264

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] U.S. EIA. 2023. Annual Energy Outlook 2023. Retrieved 2023-06-04 from https://www.eia.gov/outlooks/aeo/pdf/AEO2023_Narrative.pdf

[15] Christoph Feichtenhofer, haoqi fan, Yanghao Li, and Kaiming He. 2022. Masked Autoencoders As Spatiotemporal Learners. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 35946–35958. https://proceedings.neurips.cc/paper_files/paper/2022/file/e97d1081481a4017df96b51be31001d3-Paper-Conference.pdf

[16] D. Fox, W. Burgard, and S. Thrun. 1997. The dynamic window approach to collision avoidance. *IEEE Robotics and Automation Magazine* 4, 1 (1997), 23–33. https://doi.org/10.1109/100.580977

[17] Rui Fukushima, Kei Ota, Asako Kanezaki, Yoko Sasaki, and Yusuke Yoshiyasu. 2022. Object Memory Transformer for Object Goal Navigation. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, Philadelphia, PA, USA, 11288–11294. https://doi.org/10.1109/ICRA46639.2022.9812027

[18] Mark Fulton, Jake Baker, Margot Brandenburg, Ron Herbst, John Cleveland, Joel Rogers, and Chinwe Onyeagoro. 2012. United States Building Energy Efficiency Retrofits: Market Sizing and Financing Models. Retrieved 2023-06-20 from https://www.rockefellerfoundation.org/wp-content/uploads/United-States-Building-Energy-Efficiency-Retrofits.pdf

[19] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. 2022. CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation. arXiv:2203.10421 [cs.CV]

[20] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. 2022. Learning to Map for Active Semantic Goal Navigation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=swrMQttr6wN

[21] Heather Goetsch and Michael Deru. 2022. Operational Emissions Accounting for Commercial Buildings. Retrieved 2023-06-14 from https://www.nrel.gov/docs/fy22osti/81670.pdf

[22] Yohei Hayamizu, Saeid Amiri, Kishan Chandan, Keiki Takadama, and Shiqi Zhang. 2021. Guiding Robot Exploration in Reinforcement Learning via Automated Planning. *Proceedings of the International Conference on Automated Planning and Scheduling* 31, 1 (May 2021), 625–633. https://doi.org/10.1609/icaps.v31i1.16011

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16000–16009.

[24] Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned Language Models for Text Classification. *CoRR* abs/1801.06146 (2018). arXiv:1801.06146 http://arxiv.org/abs/1801.06146

[25] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. 2023. Visual Language Maps for Robot Navigation. arXiv:2210.05714 [cs.RO]

[26] Matan Keidar and Gal A. Kaminka. 2014. Efficient frontier detection for robot exploration. *The International Journal of Robotics Research* 33, 2 (2014), 215–236. https://doi.org/10.1177/0278364913494911 arXiv:https://doi.org/10.1177/0278364913494911

[27] Matan Keidar, Eran Sadeh-Or, and Gal A. Kaminka. 2011. Fast Frontier Detection for Robot Exploration. In *Proceedings of the 10th International Conference on Advanced Agent Technology* (Taipei, Taiwan) *(AAMAS’11)*. Springer-Verlag, Berlin, Heidelberg, 281–294. https://doi.org/10.1007/978-3-642-27216-5_20

[28] Kyungki Kim and Matthew Peavy. 2022. BIM-based semantic building world modeling for robot task planning and execution in built environments. *Automation in Construction* 138 (2022), 104247. https://doi.org/10.1016/j.autcon.2022.104247

[29] 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. 2023. Segment Anything. *arXiv:2304.02643* (2023).

[30] Constantine E. Kontokosta, Danielle Spiegel-Feld, and Sokratis Papadopoulos. 2020. The impact of mandatory energy audits on building energy use. *Nature Energy* 5, 4 (01 Apr 2020), 309–316. https://doi.org/10.1038/s41560-020-0589-6

[31] Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. 2011. G2o: A general framework for graph optimization. In *2011 IEEE International Conference on Robotics and Automation* (Shanghai, China). IEEE, 3607–3613. https://doi.org/10.1109/ICRA.2011.5979949

[32] Mathieu Labbé and François Michaud. 2019. RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. *Journal of Field Robotics* 36, 2 (2019), 416–446. https://doi.org/10.1002/rob.21831 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.21831

[33] Billy Pik Lik Lau, Brandon Jin Yang Ong, Leonard Kin Yung Loh, Ran Liu, Chau Yuen, Gim Song Soh, and U-Xuan Tan. 2022. Multi-AGV’s Temporal Memory-Based RRT Exploration in Unknown Environment. *IEEE Robotics and Automation Letters* 7, 4 (2022), 9256–9263. https://doi.org/10.1109/LRA.2022.3190628

[34] Steven M LaValle et al. 1998. Rapidly-exploring random trees: A new tool for path planning. (1998).

[35] LEED. 2014. Green Building 101: Why is energy efficiency important? Retrieved 2023-06-04 from https://www.usgbc.org/articles/green-building-101-why-energy-efficiency-important

[36] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. 2022. Language-driven Semantic Segmentation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=RriDjddCLN

[37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 12888–12900. https://proceedings.mlr.press/v162/li22n.html

[38] Hao Liu, Lisa Lee, Kimin Lee, and Pieter Abbeel. 2023. Instruction-Following Agents with Multimodal Transformer. arXiv:2210.13431 [cs.CV]

[39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv:2303.05499 [cs.CV]

[40] Mirtha Lucas, Miguel Lerma, Jacob Furst, and Daniela Raicu. 2022. RSI-Grad-CAM: Visual Explanations from Deep Networks via Riemann-Stieltjes Integrated Gradient-Based Localization. In *Advances in Visual Computing*, George Bebis, Bo Li, Angela Yao, Yang Liu, Ye Duan, Manfred Lau, Rajiv Khadka, Ana Crisan, and Remco Chang (Eds.). Springer International Publishing, Cham, 262–274.

[41] Zhenjun Ma, Paul Cooper, Daniel Daly, and Laia Ledo. 2012. Existing building retrofits: Methodology and state-of-the-art. *Energy and Buildings* 55 (2012), 889–902. https://doi.org/10.1016/j.enbuild.2012.08.018 Cool Roofs, Cool Pavements, Cool Cities, and Cool World.

[42] Andréa Macario Barros, Maugan Michel, Yoann Moline, Gwenolé Corre, and Frédérick Carrel. 2022. A Comprehensive Survey of Visual SLAM Algorithms. *Robotics* 11, 1 (2022). https://doi.org/10.3390/robotics11010024

[43] Bharadwaj R.K. Mantha, Carol C. Menassa, and Vineet R. Kamat. 2018. Robotic data collection and simulation for evaluation of building retrofit performance. *Automation in Construction* 92 (2018), 88–102. https://doi.org/10.1016/j.autcon.2018.03.026

[44] Matthew Louis Mauriello, Leyla Norooz, and Jon E. Froehlich. 2015. Understanding the Role of Thermography in Energy Auditing: Current Practices and the Potential for Automated Solutions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1993–2002. https://doi.org/10.1145/2702123.2702528

[45] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. 2018. Overcoming Exploration in Reinforcement Learning with Demonstrations. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 6292–6299. https://doi.org/10.1109/ICRA.2018.8463162

[46] Eileen Peppard. 2013. *Energy Audit Procedures*. Technical Report. University of Hawai'i. Retrieved 2023-06-20 from https://www.hnei.hawaii.edu/wp-content/uploads/Energy-Audit-Procedures.pdf

[47] Morgan Quigley, Brian Gerkey, Ken Conley, Josh Faust, Tully Foote, Jeremy Leibs, Eric Berger, Rob Wheeler, and Andrew Ng. 2009. ROS: an open-source Robot Operating System. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA) Workshop on Open Source Robotics*, Vol. 3. Kobe, Japan, 5. https://www.ros.org

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[49] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018). http://openai-assets.s3.amazonaws.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[50] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. *CoRR* abs/1606.05250 (2016). arXiv:1606.05250 http://arxiv.org/abs/1606.05250

[51] 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. 2021. 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*. https://arxiv.org/abs/2109.08238

[52] 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. 2021. 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 (Round 2)*. https://openreview.net/forum?id=-v4OuqNs5P

[53] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. 2019. Adversarial Representation Learning for Text-to-Image Matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

[54] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[55] Davide Scaramuzza and Friedrich Fraundorfer. 2011. Visual Odometry [Tutorial]. *IEEE Robotics and Automation Magazine* 18, 4 (2011), 80–92. https://doi.org/10.1109/MRA.2011.943233

[56] Dhruv Shah, Błażej Osiński, brian ichter, and Sergey Levine. 2023. LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action. In *Proceedings of The 6th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 205)*, Karen Liu, Dana Kulic, and Jeff Ichnowski (Eds.). PMLR, 492–504. https://proceedings.mlr.press/v205/shah23b.html

[57] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[58] Albert Thumann, Terry Niehus, and William Younger. 2013. *Handbook of Energy Audits* (9th ed.). River, New York. https://doi.org/10.1201/9781003151722

[59] Hassan Umari and Shayok Mukhopadhyay. 2017. Autonomous robotic exploration based on multiple rapidly-exploring randomized trees. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Vancouver, BC, Canada, 1396–1402. https://doi.org/10.1109/IROS.2017.8202319

[60] Erkan Uslu, Furkan Çakmak, Muhammet Balcılar, Attila Akıncı, M. Fatih Amasyalı, and Sırma Yavuz. 2015. Implementation of frontier-based exploration algorithm for an autonomous robot. In *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*. IEEE, Madrid, Spain, 1–7. https://doi.org/10.1109/INISTA.2015.7276723

[61] Chao Wang, Yong K. Cho, and Mengmeng Gai. 2013. As-Is 3D Thermal Modeling for Existing Building Envelopes Using a Hybrid LIDAR System. *Journal of Computing in Civil Engineering* 27, 6 (2013), 645–656. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000273

[62] Chaoqun Wang, Lili Meng, Teng Li, Clarence W. De Silva, and Max Q.-H. Meng. 2017. Towards autonomous exploration with information potential field in 3D environments. In *2017 18th International Conference on Advanced Robotics (ICAR)*. IEEE, Hong Kong, China, 340–345. https://doi.org/10.1109/ICAR.2017.8023630

[63] Cheng-Yan Wu and Huei-Yung Lin. 2019. Autonomous Mobile Robot Exploration in Unknown Indoor Environments Based on Rapidly-exploring Random Tree. In *2019 IEEE International Conference on Industrial Technology (ICIT)*. 1345–1350. https://doi.org/10.1109/ICIT.2019.8754938

[64] B. Yamauchi. 1997. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97. 'Towards New Computational Principles for Robotics and Automation'*. Association for Computing Machinery, 146–151. https://doi.org/10.1109/CIRA.1997.613851

[65] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing XU, and Hang Xu. 2022. DetCLIP: Dictionary-Enriched Visual-Concept Paralleled Pre-training for Open-world Detection. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 9125–9138. https://proceedings.neurips.cc/paper_files/paper/2022/file/3ba960559212691be13fa81d9e5e0047-Paper-Conference.pdf

[66] Jincheng Yu, Jianming Tong, Yuanfan Xu, Zhilin Xu, Haolin Dong, Tianxiang Yang, and Yu Wang. 2021. SMMR-Explore: SubMap-based Multi-Robot Exploration System with Multi-robot Multi-target Potential Field Exploration Method. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Xi'an, China, 8779–8785. https://doi.org/10.1109/ICRA48506.2021.9561328

[67] Chen Zheng, Quan Guo, and Parisa Kordjamshidi. 2020. Cross-Modality Relevance for Reasoning on Language and Vision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7642–7651. https://doi.org/10.18653/v1/2020.acl-main.683

[68] Cheng Zhu, Rong Ding, Mengxiang Lin, and Yuanyuan Wu. 2015. A 3D Frontier-Based Exploration Tool for MAVs. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, Vietri sul Mare, Italy, 348–352. https://doi.org/10.1109/ICTAI.2015.60

[69] Hongbiao Zhu, Chao Cao, Yukun Xia, Sebastian Scherer, Ji Zhang, and Weidong Wang. 2021. DSVP: Dual-Stage Viewpoint Planner for Rapid Exploration by Dynamic Expansion. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Prague, Czech Republic, 7623–7630. https://doi.org/10.1109/IROS51168.2021.9636473