Mobile Robot Place Segmentation and Categorization Using Object Semantics

Jesus Moncada-Ramirez[1], Jose-Raul Ruiz-Sarmiento[1], Cipriano Galindo[1] and Javier Gonzalez-Jimenez[1]

Abstract— **A mobile robot's ability to perform complex tasks depends critically on its understanding of the environment in which it operates. This understanding—often represented as a map—must incorporate both semantic information about objects (identity, functionality, ...) and a high-level grasp of spatial organization. This latter capability, referred to as place segmentation and categorization, is essential for reasoning in diverse and dynamic spaces. However, conventional semantic mapping methods typically segment environments into predefined rooms and rely on closed-set labels, limiting adaptability and expressiveness in real-world, multi-functional settings.**

In this paper, we propose a novel pipeline for indoor place segmentation and categorization that leverages rich object semantics to generate more flexible and meaningful spatial representations. Rather than adhering to rigid architectural boundaries, our approach defines places as clusters of functionally coherent objects. These clusters are formed using descriptors that integrate geometric location with semantic information, derived from context-aware functional descriptions of objects generated by Large Language Models (LLMs) and encoded as sentence embeddings. We apply density-based clustering to identify semantically meaningful places and use LLMs to assign each cluster open-set natural-language tags and detailed functional descriptions, resulting in enriched and interpretable maps.

Experimental results on the ScanNet and SceneNN datasets demonstrate that our method significantly improves segmentation consistency over geometric and basic semantic baselines while producing coherent and informative place categorizations that better reflect the functional layout of indoor environments.

I. INTRODUCTION

Mobile robots are increasingly expected to perform complex tasks across a wide range of domains, including home assistance, industrial automation, healthcare, and education. To meet these demands, robots must possess advanced cognitive capabilities that allow them to interpret their surroundings and interact with them purposefully and intelligently.

Central to these capabilities is the availability of semantically rich representations of the world—commonly referred to as semantic maps [1], [2]—alongside a reasoning and execution engine capable of effectively leveraging this knowledge [3]. Semantic maps integrate information about both the spatial structure of the environment (e.g., objects, regions, boundaries) and the associated semantic content (e.g., attributes, functionalities, inter-object relationships), providing a foundation for deeper environmental understanding and goal-directed behavior of robots. Ultimately, the richness of these semantic representations plays a crucial role in determining the variety of tasks that mobile robots can perform effectively within their environments.

An effective way to improve the richness of such representations is to include information about the different areas or places into which the environment is divided. This process usually involves segmenting the workspace into meaningful regions or places, known as *place segmentation*, followed by assigning semantic labels or descriptions to each area, a process referred to as *place categorization*. For example, in a supermarket environment, such a map could include a dairy section (place), which serves the purpose of offering refrigerated food products (functionality) and contains objects such as refrigerators and shelves (relationships). The map could also encode that refrigerators are used to store perishable items, like milk and cheese (functionality), and have a tall, rectangular shape (property); or that shopping carts and checkout counters are used together during the purchasing process (relationships and functionality).

Traditionally, approaches to building semantic representations that include place information have faced some limitations. First, some methods rely on manual annotation, a process that is time-consuming, prone to human errors, and typically results in oversimplified representations unable to capture the complexity of real-world environments. Second, these approaches are commonly tied to the concept of *rooms*, framing the task as building a map and segmenting it into different rooms. This leads to semantic inconsistencies, particularly in large spaces where multiple activities—and thus diverse objects—may coexist. In such cases, assigning a single room label limits the information given to the robot, which may affect its performance. Lastly, these approaches operate under the *closed-set* assumption, which means that the system can only make use of information that is either included in a predefined repository (such as an ontology) or encountered during the training stage of a categorization model, ignoring any new or unseen concepts during operation.

The emergence of Deep Learning (DL) techniques, such as text or image embedding models [4]–[6], and generative Artificial Intelligence (AI) models, like Large Language Models (LLMs) [7], opens the door to exploring new approaches to solve classical problems. In this work, we address the limitations mentioned above by exploiting advanced AI models in a clustering-based pipeline for place segmentation and categorization. Our method operates on a pre-built object-oriented semantic map using both objects' spatial locations in the environment and their semantics—obtained by different Natural Language Processing (NLP) embedding models—to

[1] Machine Perception and Intelligent Robotics Group (MAPIR-UMA), Malaga Institute for Mechatronics Engineering and Cyber-Physical Systems (IMECH.UMA), University of Malaga, Spain

build object descriptors. These descriptors are then processed by a classical clustering algorithm that groups them into meaningful places, which are not restricted to conventional rooms. After grouping objects into different geometrically and semantically coherent clusters, our method categorizes each place by exploiting the vast knowledge encoded in LLMs, which results in an open set of place possibilities. To validate our proposal, we also contribute a places-annotated dataset that permits ablation studies. Experimental results show that including object semantics significantly enhances place segmentation performance. Moreover, the use of LLMs for place categorization produces coherent place descriptions. Both the dataset and the implementation used in our experiments are publicly available at `https://github.com/MAPIRlab/generative-topological-maps`.

II. Related Work

Classical methods for dealing with place segmentation and categorization can be grouped according to the type of input data. Some methods rely on geometric information to exploit the structural layout of the environment [8], [9]. The geometric characteristics of the different areas, i.e., rooms and corridors, are used to segment and categorize the robot workspace.

Other studies improve place categorization by leveraging the identification of objects and their relationships in the scene [1], [10]. These methods are also bound to the room concept and consider only a limited number of object categories and relationships (closed-set assumption), reducing their ability to face diverse scenarios.

The closed-set issue is mitigated by methods that directly infer the place categorization from a given image, using DL [11]–[13]. This technique offers limited open-set capabilities, depending on the size of the training data, but still lacks the benefits of using LLMs to provide richer place categorizations and descriptions.

Modern semantic mapping pipelines, such as Kimera [14] and Hydra [15], combine dense 3D SLAM with semantic representations of the environment. These systems construct hierarchical maps that integrate geometric and semantic data, enabling richer spatial understanding. However, they remain bound to the room concept, which may limit the robot's ability to reason about open and multifunctional spaces beyond domestic environments.

Our work abandons this *room-centric* perspective by relying on the semantic relationships of objects to segment and categorize the environment. The result enhances the semantic map with a division of the space, e.g., a room, into different areas according to their functionality. In this sense, our work can be considered as an affordance-based approach. Some works have also explored affordances but in different directions, like path-planning [16], [17] or grasping [18], [19].

III. Problem Formulation

In essence, a semantic map $\mathbf{m}$ of a 3D environment with N objects can be defined as follows:

$$\mathbf{m} = \{\mathbf{o}_1, \ldots, \mathbf{o}_N \mid \mathbf{o}_i = (c_i, d_i, l_i, \mathbf{p}_i^{\text{object}}), \forall i \in \{1, \ldots, N\}\}$$

where:

- $c_i \in \mathbb{R}^3$ represents the i-th object's location (e.g., the bounding box centroid).
- $d_i \in \mathbb{R}^3$ represents the i-th object's dimensions (e.g., the bounding box size).
- $l_i \in \mathcal{L}$ is the semantic label assigned to the i-th object, with $\mathcal{L}$ being the set of all possible semantic classes (e.g., table, chair, vase, etc.).
- $\mathbf{p}_i^{\text{object}}$ is a set of additional properties describing the i-th object.

Note that $\mathbf{p}_i^{\text{object}}$ can be used to store supplementary information relevant to the problem at hand, such as uncertainties in object classifications, multiple semantic labels each with a confidence score, or even relationships with other objects in the map.

In this context, the *place segmentation* task consists of, given a pre-built semantic map of a scene, organizing objects into distinct subsets—referred to as places—containing objects from the map that are both geometrically and semantically related. Formally, given $\mathbf{m}$, a place segmentation function ψ produces a set of places $\mathcal{P}$:

$$\mathcal{P} = \psi(\mathbf{m}) = \{P_k \mid k \in K\}$$

where K is the set of place indices, and each place P_k consists of a subset of the objects in the map:

$$P_k = \{\mathbf{o}_j \mid j \in J_k \subseteq \{1, \ldots, N\}\}$$

Note that J_k is the set of indices of objects in the k-th place, and $\{1, \ldots, N\}$ is the set of all object indices.

Once the set of places with their objects has been defined, each place can be enhanced with additional semantic information, thereby improving the semantic map's usability. This step is commonly referred to as *place categorization*. One way to do this is by generating descriptions based on the objects contained in each place. Formally, given a place-segmented map $\mathcal{P} = \{P_k \mid k \in K\}$, a place categorization function ϕ assigns to each place P_k a set of semantic properties $\mathbf{p}_k^{\text{place}}$:

$$\phi(P_k) = \mathbf{p}_k^{\text{place}}$$

These properties may include a tag $t_k \in \mathbf{p}_k^{\text{place}}$ used to identify the place in a general way, or a natural-language description $d_k \in \mathbf{p}_k^{\text{place}}$ including a more detailed explanation.

The goal of a place segmentation function ψ is to generate a place segmentation $\mathcal{P}$, ensuring that objects within every place P_k are geometrically and semantically consistent while maintaining low similarities among objects in different groups. In turn, the goal of the place categorization function ϕ is to assign to each place a set of semantic characteristics $\mathbf{p}_k^{\text{place}}$ that best describes the corresponding region of the map.

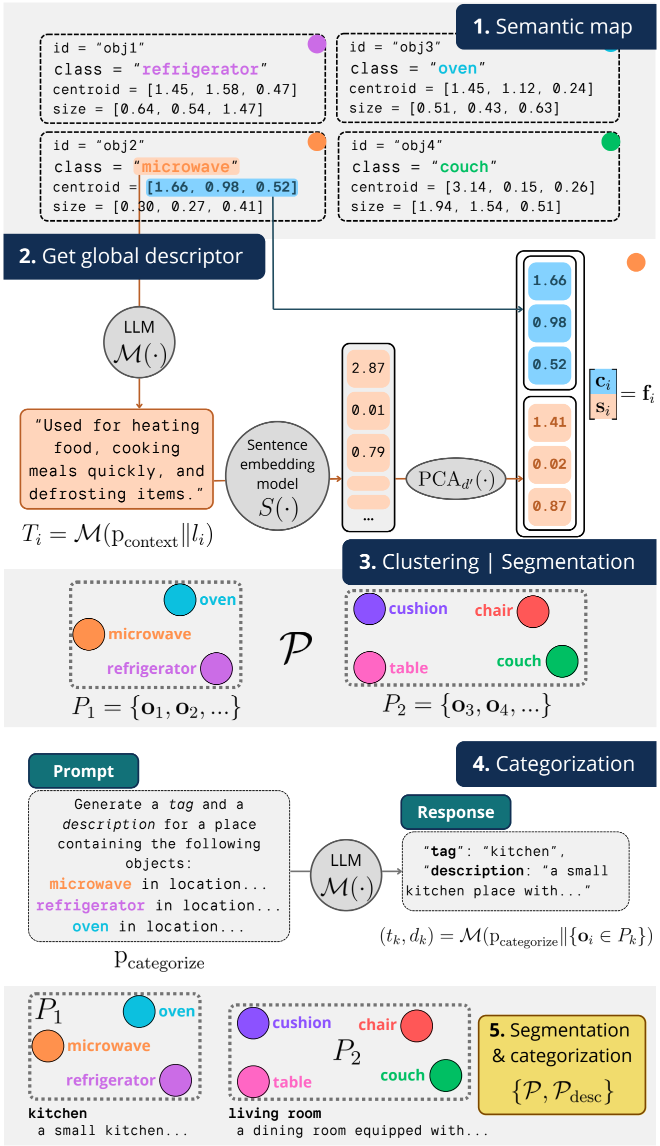

1. Semantic map
id = "obj1"
class = "refrigerator"
centroid = [1.45, 1.58, 0.47]
size = [0.64, 0.54, 1.47]
id = "obj3"
class = "oven"
centroid = [1.45, 1.12, 0.24]
size = [0.51, 0.43, 0.63]
id = "obj2"
class = "microwave"
centroid = [1.66, 0.98, 0.52]
size = [0.30, 0.27, 0.41]
id = "obj4"
class = "couch"
centroid = [3.14, 0.15, 0.26]
size = [1.94, 1.54, 0.51]
2. Get global descriptor
LLM
$\mathcal{M}(\cdot)$
"Used for heating food, cooking meals quickly, and defrosting items."
$T_i = \mathcal{M}(\mathrm{p}_{\text{context}} \| l_i)$
Sentence embedding model
$S(\cdot)$
2.87
0.01
0.79
...
$\mathrm{PCA}_{d'}(\cdot)$
1.66
0.98
0.52
1.41
0.02
0.87
$\begin{bmatrix}\mathbf{c}_i\\ \mathbf{s}_i\end{bmatrix} = \mathbf{f}_i$
3. Clustering | Segmentation
oven
microwave
refrigerator
$P_1 = \{\mathbf{o}_1, \mathbf{o}_2, ...\}$
$\mathcal{P}$
cushion
chair
table
couch
$P_2 = \{\mathbf{o}_3, \mathbf{o}_4, ...\}$
4. Categorization
Prompt
Generate a tag and a description for a place containing the following objects:
microwave in location...
refrigerator in location...
oven in location...
$\mathrm{p}_{\text{categorize}}$
LLM
$\mathcal{M}(\cdot)$
Response
"tag": "kitchen",
"description: "a small kitchen place with..."
$(t_k, d_k) = \mathcal{M}(\mathrm{p}_{\text{categorize}} \| \{\mathbf{o}_i \in P_k\})$
P_1
oven
microwave
refrigerator
kitchen
a small kitchen...
cushion
chair
P_2
table
couch
living room
a dining room equipped with...
5. Segmentation & categorization
$\{\mathcal{P}, \mathcal{P}_{\text{desc}}\}$

Fig. 1: Overview of the presented place segmentation and categorization pipeline.

IV. METHOD

The problem formulated in Section III can be addressed by applying clustering to a set of object descriptors such that each resulting cluster corresponds to a distinct place. A straightforward approach is to use only the spatial position of objects as descriptors. However, this may group spatially close but semantically unrelated objects into the same place. Semantic information, primarily derived from the object class, can be included in the descriptor to overcome this issue (see Section IV-A).

Once places on the map are defined, LLMs can be used to categorize them. Specifically, LLMs can generate semantic information, such as a short tag and a natural-language description, for each place by processing a prompt including the set of objects within it. This prompt can be phrased as: "Describe a place containing the following objects: [...]" (see Section IV-B). An overview of the proposed place segmentation and categorization pipeline is shown in Fig. 1.

A. *Geometric-semantic clustering*

To address the presented problem, we propose a method that applies clustering to a set of object descriptors containing both geometric and semantic information. This way, objects are mapped to a feature space where proximity means both spatial closeness and semantic similarity.

Formally, for each object in the semantic map $\mathbf{o}_i \in \mathbf{m}$, we define a descriptor vector $\mathbf{f}_i \in \mathbb{R}^{3+d'}$ composed of two parts:

$$\mathbf{f}_i = \begin{bmatrix}\mathbf{c}_i\\ \mathbf{s}_i\end{bmatrix}$$

where:

- $\mathbf{c}_i \in \mathbb{R}^3$ is the object's geometric descriptor, e.g., the bounding box centroid.
- $\mathbf{s}_i \in \mathbb{R}^{d'}$ is the object's semantic descriptor.

The semantic descriptor $\mathbf{s}_i$ can be obtained in several ways and defines how the semantic information is used in our context. A simple approach consists of applying a pre-trained word embedding function $W : \mathcal{L} \to \mathbb{R}^d$ (e.g., BERT [4] or RoBERTa [5]) to the object's semantic label $l_i \in \mathcal{L}$:

$$\mathbf{s}_i = \mathrm{PCA}_{d'}(W(l_i))$$

where $\mathrm{PCA}_{d'}(\cdot)$ denotes the projection onto $\mathbb{R}^{d'}$ via Principal Component Analysis (PCA), which reduces the dimensionality of the embedding.

However, this approach implicitly prioritizes the semantics encoded in the word embedding function, which may not be optimal for our specific context. For example, the embedding function $W(\cdot)$ may consider semantic labels such as "cashbox" and "fridge" to be semantically similar because they are electronic devices. However, in the context of place segmentation, these objects do not share any functionality and typically belong to distinct places; thus, they should not be close in the semantic space.

To address this limitation, we propose generating a context-aware description T_i for each object $\mathbf{o}_i$ using an LLM. Specifically, the LLM $\mathcal{M}$ is provided with a prompt $\mathrm{p}_{\text{context}}$ that includes the object's semantic label l_i and requests a short description tailored for the task at hand, i.e., including contextual information about the functionality the object typically serves:

$$T_i = \mathcal{M}(\mathrm{p}_{\text{context}} \| l_i)$$

Note that $\|$ denotes concatenation. The semantic descriptor is then obtained by applying a sentence embedding model S : Text $\to$ $\mathbb{R}^d$ (e.g., Sentence-BERT [6]) to the contextualized sentence T_i:

$$\mathbf{s}_i = \mathrm{PCA}_{d'}(S(T_i))$$

where PCA is used again to reduce dimensionality. This way, the semantics of the object's class are not encoded according to the word embedding model, but rather to the functionality it presents in our context.

Once each object descriptor is calculated, it is important to independently normalize the geometric and semantic components to ensure a consistent scaling in both feature types. With the descriptors normalized, we can perform clustering to group objects into places. In this work, we opt for using DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) [20] because it allows the detection of clusters with arbitrary shape and size, and can automatically identify low-density regions as noise, which is particularly useful in semantic maps with complex structures. Moreover, since it does not require the number of clusters as input, DBSCAN naturally adapts to scenarios where the number of places is unknown a priori.

To apply clustering, a distance metric between descriptors must be defined. In this work, we propose a weighted distance function that gives more importance to spatial proximity over semantic similarity. Otherwise, the clustering process may tend to group objects with the same functionality in the same cluster, even if they are on opposite sides of the space we are segmenting. Given two object descriptors $\mathbf{f}_i = [\mathbf{c}_i^\top, \mathbf{s}_i^\top]^\top$ and $\mathbf{f}_j = [\mathbf{c}_j^\top, \mathbf{s}_j^\top]^\top$, the custom distance function $d_{\text{cluster}}(\mathbf{f}_i, \mathbf{f}_j)$ used in DBSCAN is defined as follows:

$$d_{\text{cluster}}(\mathbf{f}_i, \mathbf{f}_j) = \|\mathbf{c}_i - \mathbf{c}_j\|_2 + \alpha_s \cdot \frac{\|\mathbf{s}_i - \mathbf{s}_j\|_2}{\sqrt{d'}}$$

where:

- α_s is the semantic weight.
- d' is the dimension of the semantic descriptor after applying the dimensionality reduction, if any.

Note that DBSCAN may label some objects as noise, meaning they do not belong to any cluster due to low local density. In our approach, these are treated as individual clusters to preserve their uniqueness.

B. LLM categorization

After the execution of the previous stages, each resulting cluster $P_k \in \mathcal{P}$ contains a subset of objects and defines a place on the map. To enhance the richness of these places, a categorization stage can be performed, assigning a short tag and a natural-language description to each place.

To achieve this, we leverage the natural language understanding capabilities of LLMs. Specifically, we build a prompt $\mathrm{p}_{\text{categorize}}$ where all the information regarding each cluster's objects is included. This prompt's instruction requests the model to generate both the tag t_k and the description d_k, by attending to the objects present in the place. Formally, being $\mathcal{M}$ the LLM, for each cluster P_k we perform:

$$(t_k, d_k) = \mathcal{M}(\mathrm{p}_{\text{categorize}} \| \{\mathbf{o}_i \in P_k\})$$

The prompt $\mathrm{p}_{\text{categorize}}$ is a prompt template, containing placeholders where the information regarding the problem is inserted. Simply put, it can be phrased as "Describe a place containing the following objects: [...]". The exact prompt template used in our implementation is available in the public project repository (see *README* file).

V. EVALUATION

To quantitatively evaluate the place segmentation pipeline and analyze several alternatives in ablation studies, we conducted experiments using a self-created dataset (see Section V-A), given the peculiarities introduced by the concept of place. The place categorization part was evaluated qualitatively. The implementation relies exclusively on open-source technologies, including word and sentence embedding models and LLMs; the chosen clustering algorithm was DBSCAN. For the quantitative evaluation of the place segmentation part of the pipeline, three clustering metrics were used (see Section V-B). Results show that incorporating semantic information improves place segmentation performance. The place categorization stage generates coherent descriptions of places (see Section V-C).

A. Dataset

The dataset used for evaluation consists of multiple ground-truth place segmentations and categorizations for each of the ten semantic maps selected from the publicly available and widely used ScanNet [21] and SceneNN [22] datasets. For each semantic map $\mathbf{m}^{(j)}$, we provide a set of valid place segmentations $\left\{\mathcal{P}_{\text{GT}}^{(j,\ell)}\right\}_{\ell=1}^{s_j}$ where each $\mathcal{P}_{\text{GT}}^{(j,\ell)}$ is defined as:

$$\mathcal{P}_{\text{GT}}^{(j,\ell)} = \left\{(P_k, t_k, d_k) \mid k \in K^{(j,\ell)}\right\}$$

Here, P_k is a set of objects grouped as a place, t_k is a short tag, and d_k is a textual description. All maps and segmentations are stored in JSON format. For each semantic map, a set of place segmentation results is provided instead of a single one, since multiple valid segmentations of the objects may exist.

From the ScanNet dataset, we selected large scenes containing a high number of objects, specifically from five different one-story apartments: `scene_0000`, `scene_0101`, `scene_0392`, `scene_0515`, and `scene_0673`. Scenes from the SceneNN dataset were generally smaller, contained fewer objects, and included a lounge (scene `011`), offices (scenes `030`, `078`, `086`), and a bedroom (scene `096`).

B. Evaluation setup

All models were obtained from the Hugging Face Model Hub[1]. The word embedding models used in the experiments were BERT (`bert-base-uncased`) [4] and RoBERTa (`roberta-base`) [5]. The sentence embedding model employed was `sentence-transformers/all-mpnet-base-v2`. Although this model is not exactly Sentence-BERT [6], it belongs to the Sentence Transformers family, following the same training methodology, and is widely used for sentence-level embedding tasks. The LLM used was DeepSeek Qwen 14B (`deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`). The semantic weight employed in the clustering distance was empirically set to $\alpha_s = 0.55$. Semantic

[1] `https://huggingface.co/models`

Method	NMI	V-Measure	ARI
LLM + Sentence emb.	**0.611**	**0.586**	**0.314**
Word emb. (BERT)	0.584	0.560	0.272
Word emb. (RoBERTa)	0.531	0.507	0.201
Geometric	0.351	0.330	0.210

TABLE I: Mean performance of place segmentation alternatives ordered by NMI.

embeddings were reduced to $d' = 3$ using PCA, as this dimensionality maintained clustering performance and reduced computational cost. For clustering, we ran DBSCAN with `min_samples` set to 2 and `epsilon` varying from 1.0 to 2.0 in 0.1 increments. To decouple our evaluation from any single parameter choice, we averaged each clustering metric across all tested `epsilon` values and reported these mean scores.

To quantitatively evaluate the place segmentation performance against the ground truth, we employ three different clustering evaluation metrics, each capturing different aspects in our context:

- Normalized Mutual Information (NMI) [23] quantifies the amount of shared information between two clustering results, i.e., how the segmented places preserve the information in the ground truth.
- V-Measure [24] measures homogeneity (each predicted place contains only objects belonging to a single ground truth place) and completeness (all objects belonging to the same ground truth place are assigned to the same predicted place).
- Adjusted Rand Index (ARI) [25] evaluates how accurately the system assigns objects to the same place as in the ground truth, compared to a random assignment.

For semantic maps including multiple ground truth place segmentations, each method was evaluated against all available options, and the best score obtained was considered.

C. Results

An evaluation of each method's place segmentation performance against the ground truth is presented in Table I.

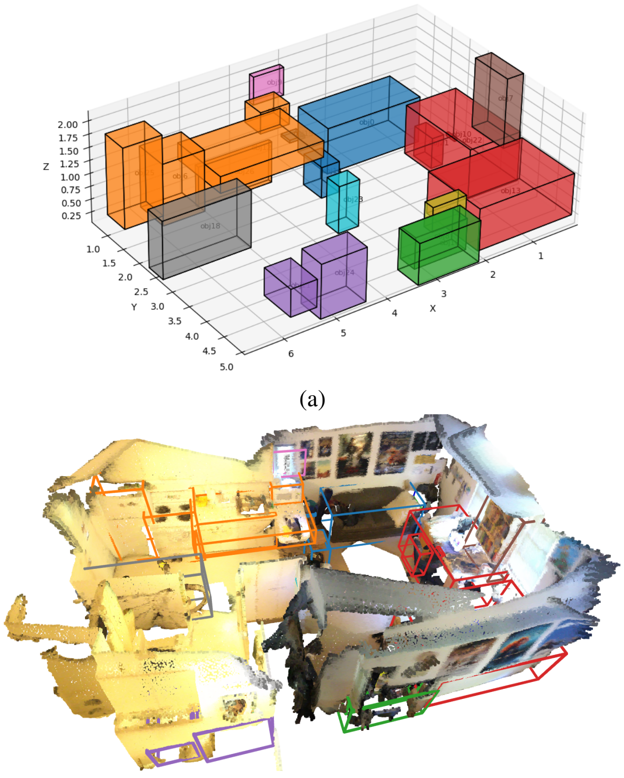
Based on the results, the core idea of clustering object descriptors enriched with semantic information leads to substantial improvements in place segmentation performance across all clustering-based metrics. Among the evaluated semantic descriptors, the pipeline where an LLM generates a contextualized sentence of each object's functionality and a sentence embedding model encodes it achieves the highest NMI (0.611), V-Measure (0.586), and ARI (0.314). On the other hand, purely geometric place segmentation performs significantly worse compared to all the semantic models.

The other semantic descriptors still provide promising results, though they fall short of the complete pipeline. These results demonstrate the importance of contextualizing the generated embeddings according to the task at hand, rather than relying solely on the inherent encoding from embedding models.

Dataset	Method	NMI	V-Measure	ARI
ScanNet [21]	LLM + Sentence emb.	**0.60**	**0.58**	**0.27**
	Word emb. (BERT)	0.56	0.54	0.22
	Word emb. (RoBERTa)	0.51	0.49	0.12
	Geometric	0.27	0.25	0.12
SceneNN [22]	LLM + Sentence emb.	**0.62**	**0.59**	**0.35**
	Word emb. (BERT)	0.61	0.59	0.33
	Word emb. (RoBERTa)	0.55	0.52	0.28
	Geometric	0.43	0.41	0.30

TABLE II: Mean performance of place segmentation alternatives grouped by dataset, ordered by NMI.

Results grouped by dataset are presented in Table II. As shown, all methods perform slightly better on the SceneNN dataset than on ScanNet. This may be because SceneNN scenes are more compact and contain objects already grouped by functionality, making clustering easier. Conversely, ScanNet's environments—which are larger, contain more diverse objects, and present more complex spatial distributions—introduce greater semantic and geometric variability, making clustering more challenging. The ranking of approaches by clustering performance remains consistent across both datasets and aligns with the global results in Table I.



(a)

(b)

Fig. 2: Segmentation result for ScanNet scene `scene0101_00`, showing the abstracted 3D representation with colored bounding boxes (a), and the same segmentation overlaid on the point cloud of the scene (b). Each color corresponds to a different place.

An illustrative example of a place segmentation produced by the proposed pipeline is shown in Fig. 2. The result demonstrates that the method produces coherent place segmentations, grouping objects not only based on geometric proximity but also according to their functional relationships. For instance, the objects highlighted in orange in Fig. 2 have been assigned to the same place, as they all are close and correspond to kitchen-related items. This cluster includes `obj6` (stove), `obj12` (countertop), and `obj19` (microwave), among others. The method also handles outlier cases effectively. For example, the gray cluster contains only `obj18` (bicycle), which is segmented separately due to its semantic dissimilarity with nearby places (the kitchen place in orange and the bathroom place in purple).

An example of the place categorization stage is shown for the mentioned orange cluster, where the method produced the following short tag: `kitchen area` and description:

```
This location features essential kitchen
appliances such as a stove, microwave, and
refrigerator, along with a countertop and
storage cabinet, indicating a space designed
for food preparation and cooking.
```

As shown, the LLM successfully generates both the tag and the description by analyzing the set of objects included in the place, increasing the richness of the semantic map.

VI. CONCLUSIONS

In this work, we introduced an improved method for semantic map generation, focusing on meaningful place segmentation and categorization. Our technique clusters objects using combined geometric and semantic descriptors, avoiding the rigidity of predefined rooms. We generate these semantic descriptors using Large Language Models (LLMs) and sentence embeddings to capture contextual object function. Furthermore, LLMs are employed to categorize the resulting places with open-set, natural-language descriptions based on object content, eliminating closed-set constraints.

We validated our segmentation pipeline using a newly contributed dataset composed of ScanNet and SceneNN maps annotated with functional place ground truth. The results demonstrate the value of our approach: including contextual semantic information significantly enhances place segmentation accuracy and produces richer, more descriptive maps. This is especially crucial in environments where geometric information alone proves ambiguous for place determination, highlighting the effectiveness of our semantic-driven method for complex robot operation.

As future work, we plan to validate the proposed pipeline in non-domestic environments, where we believe it has even greater potential. In addition, post-processing existing clusters based purely on semantic information could improve map quality.

ACKNOWLEDGMENT

This work has been developed in the context of the project Voxeland (JA.B1-09) funded by the University of Malaga, and the project MINDMAPS (PID2023-148191NB-I00) funded by the Ministry of Science, Innovation, and Universities of Spain.

REFERENCES

[1] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J.-A. Fernandez-Madrigal, and J. Gonzalez-Jimenez, “Multi-hierarchical semantic maps for mobile robotics,” in *IROS*, 2005, pp. 2278–2283.

[2] J.-R. Ruiz-Sarmiento, C. Galindo, and J. Gonzalez-Jimenez, “Building multiversal semantic maps for mobile robot operation,” *Knowledge-Based Systems*, vol. 119, pp. 257–272, 2017.

[3] J. Moncada-Ramirez, J.-L. Matez-Bandera, J. Gonzalez-Jimenez, and J.-R. Ruiz-Sarmiento, “Agentic workflows for improving large language model reasoning in robotic object-centered planning,” *Robotics*, vol. 14, no. 3, 2025.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT*, vol. 1, 2019, pp. 4171–4186.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[6] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *EMNLP-IJCNLP*, 2019, pp. 3982–3992.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NIPS*, 2017, pp. 5998–6008.

[8] A. Pronobis and P. Jensfelt, “Large-scale semantic mapping and reasoning with heterogeneous modalities,” in *ICRA*, 2012, pp. 3515–3522.

[9] O. M. Mozos, C. Stachniss, and W. Burgard, “Supervised learning of places from range data using adaboost,” in *ICRA*, 2005, pp. 1730–1735.

[10] S. Vasudevan and R. Siegwart, “Bayesian space conceptualization and place classification for semantic maps in mobile robotics,” *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 522–537, 2008.

[11] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, “On the performance of convnets in place recognition,” in *IROS*, 2015, pp. 4297–4304.

[12] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, “Multi-modal semantic place classification,” *IJRR*, vol. 29, no. 2–3, pp. 298–320, 2010.

[13] M. Mancini, L. Porzi, S. R. Bulò, B. Caputo, and E. Ricci, “Boosting domain adaptation by discovering latent domains,” in *CVPR*, 2018, pp. 3771–3780.

[14] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, “Kimera: An open-source library for real-time metric-semantic localization and mapping,” in *ICRA*, 2020, pp. 1689–1696.

[15] Y. C. Nathan Hughes and L. Carlone, “Hydra: A real-time spatial perception system for 3d scene graph construction and optimization,” in *RSS*, 2022.

[16] F. Bark, K. Daun, and O. von Stryk, “Affordance-based actionable semantic mapping and planning for mobile rescue robots,” in *SSRR*, 2023, pp. 53–60.

[17] B. Bolte, A. Wang, J. Yang, M. Mukadam, M. Kalakrishnan, and C. Paxton, “Usa-net: Unified semantic and affordance representations for robot memory,” in *IROS*, 2023, pp. 1–8.

[18] T.-T. Do, A. Nguyen, and I. Reid, “Affordancenet: An end-to-end deep learning approach for object affordance detection,” in *ICRA*, 2018.

[19] C. Xu, Y. Chen, H. Wang, S.-C. Zhu, Y. Zhu, and S. Huang, “Partafford: Part-level affordance discovery from 3d objects,” *arXiv preprint arXiv:2202.13519*, 2022.

[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *KDD*, 1996, pp. 226–231.

[21] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “Scannet: Richly-annotated 3d reconstructions of indoor scenes,” in *CVPR*, 2017, pp. 2432–2443.

[22] B.-S. Hua, Q.-H. Pham, D. T. Nguyen, M.-K. Tran, L.-F. Yu, and S.-K. Yeung, “Scenenn: A scene meshes dataset with annotations,” in *3DV*, 2016, pp. 92–101.

[23] N. X. Vinh, J. Epps, and J. Bailey, “Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance,” *JMLR*, vol. 11, no. 95, pp. 2837–2854, 2010.

[24] A. Rosenberg and J. Hirschberg, “V-measure: A conditional entropy-based external cluster evaluation measure,” in *EMNLP-CoNLL*, 2007, pp. 410–420.

[25] L. Hubert and P. Arabie, “Comparing partitions,” *Jorunal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.