Mapping the Unseen: Unified Promptable Panoptic Mapping with Dynamic Labeling using Foundation Models

Mohamad Al Mdfaa[1], Raghad Salameh[2], Sergey Zagoruyko[1] and Gonzalo Ferrer[1]

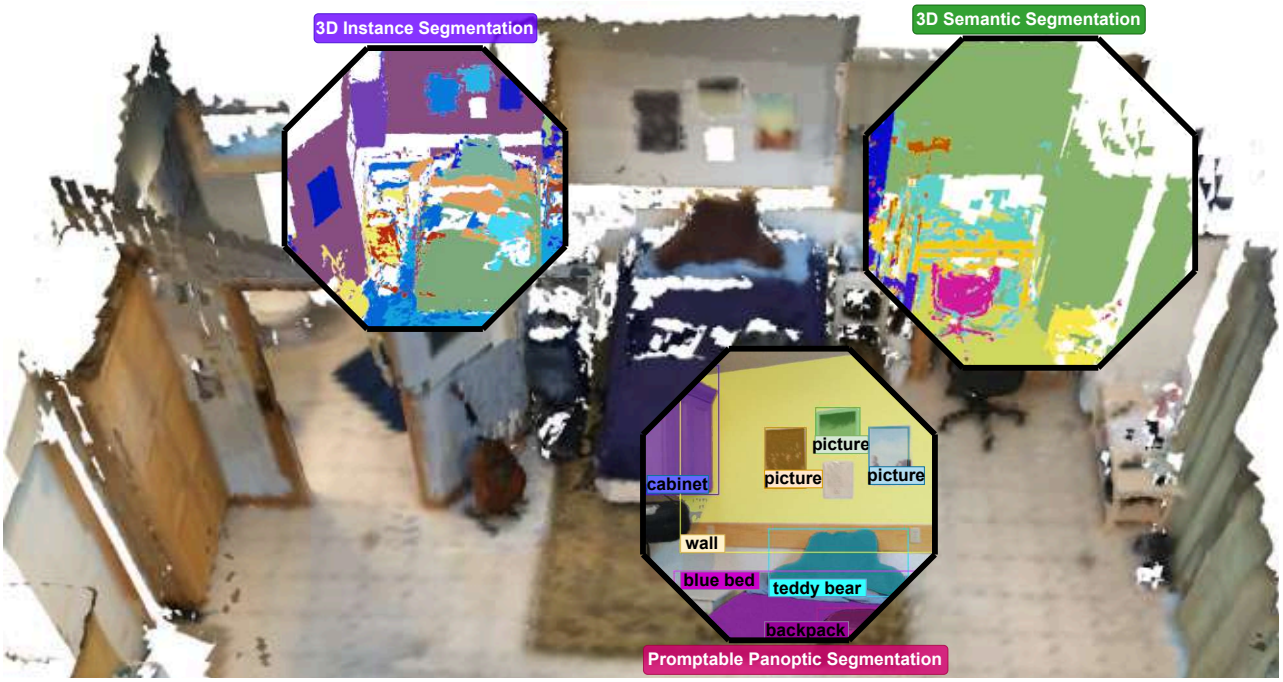


Fig. 1: Our Unified Promptable Panoptic Mapping (UPPM) dynamically generates rich object labels through open-vocabulary prompts, merging diverse yet imperfect labels into a unified semantic structure. Efficient postprocessing, coupled with open-vocabulary object detection, ensures accurate 2D segmentations and reconstructs a multi-resolution multi-TSDF map. This map enables natural human-robot communication and versatile applications, showcasing scene exploration and object retrieval using intuitive prompts. (Webpage)

Abstract*—In the field of robotics and computer vision, efficient and accurate semantic mapping remains a significant challenge due to the growing demand for intelligent machines that can comprehend and interact with complex environments. Conventional panoptic mapping methods, however, are limited by predefined semantic classes, thus making them ineffective for handling novel or unforeseen objects. In response to this limitation, we introduce the Unified Promptable Panoptic Mapping (UPPM) method. UPPM utilizes recent advances in foundation models to enable real-time, on-demand label generation using natural language prompts. By incorporating a dynamic labeling strategy into traditional panoptic mapping techniques, UPPM provides significant improvements in adaptability and versatility while maintaining high performance levels in map reconstruction. We demonstrate our approach on real-world and simulated datasets. Results show that UPPM can accurately reconstruct scenes and segment objects while generating rich semantic labels through natural language interactions. A series of ablation experiments validated the advantages of foundation model-based labeling over fixed label sets.

[1]The authors with the Skolkovo Institute of Science and Technology (Skoltech), Center for AI Technology.{mohamad.almdfaa, s.zagoruyko,g.ferrer}@skoltech.ru
[2]raghadsalameh1@gmail.com

I. Introduction

Panoptic mapping is a key element in enabling machines to comprehend and reconstruct their surroundings with advanced understanding. The representation of rich semantic information forms the bedrock for intelligent machine perception. However, prevalent methodologies [1]–[3] encounter challenges due to their reliance on rigid predefined class sets, constraining their adaptability to unforeseen objects or dynamic contexts. The applicability of these systems in diverse real-world settings necessitates either a substantial corpus of labeled data [4], [5] or the imposition of controlled environmental conditions. Integrating open-set approaches [6], [7] presents a promising avenue for enhancing robustness and generalizability in perception tasks.

This work proposes a prompt-based panoptic mapping pipeline that leverages recent advances in foundation models to enable **on-demand label generation** through natural language interactions. This approach overcomes the

limitations of static labels, allowing robots to dynamically acquire and apply object labels, resulting in **richer semantic understanding** and increased flexibility. We define **Dynamic Labeling** as the process of assigning *semantically unified categories* (see fig. 4) to the *detected objects* in previously *unseen environments*, all while preserving the *rich labels* generated through *open-vocabulary methods*. Extensive evaluations on real-world and simulated datasets demonstrate the accuracy of scene reconstruction and the effectiveness of dynamic label generation. This research opens doors for more natural human-robot communication and adaptable machine perception in dynamic environments.

II. Related Work

Semantic mapping and visual SLAM have been active areas of research, with methods proposed for dense semantic mapping [8], [9], object-centric mapping [2], [3], as well as Keypoint-based Object-level SLAM [10]. These approaches have contributed significantly to understanding complex real-world environments by focusing on object-level semantic mapping.

Dense semantic mapping methods like SemanticFusion [8] and DA-RNN [9] assign semantic labels to map elements like voxels or surfels. While enabling comprehensive scene understanding, these methods face challenges in distinguishing individual objects within the scene.

In contrast, object-centric approaches such as SLAM++ [3] and Fusion++ [2] have shown a strong focus on reconstructing specific objects, yet they often lack the ability to incorporate the semantics and geometry of background regions, limiting their capacity for a holistic understanding of scenes.

Moreover, recent advancements in panoptic mapping have aimed at addressing the limitations of traditional methods. The work on Panoptic Multi-TSDFs [11] and Panoptic Fusion [1] have significantly contributed to the field by enabling flexible representations for online multi-resolution volumetric mapping with a focus on long-term dynamic scene consistency and semantic understanding at different levels of granularity.

These panoptic approaches have shown promise in simultaneously capturing both semantic information and geometric details, providing a more comprehensive understanding of scenes. However, these methods are often constrained by predefined sets of semantic classes, limiting their adaptability to unforeseen objects and scenarios.

Our system is inspired by the recent success of generative AI and foundation models [12]–[15] in wide spectrum of tasks, such as, Image Captioning [16], [17], Object Detection [18], [19], Image Segmentation [12], etc.

The proposed UPPM aims to address these limitations by leveraging foundation models for dynamic label generation. UPPM offers a more adaptive and versatile labeling mechanism by enabling on-demand, dynamic label generation for mapped objects through natural language prompts. This new approach significantly enhances the adaptability and robustness of semantic mapping systems, fostering interactive robotic perception and human-robot communication.

III. Methodology

UPPM proposes a novel solution to address the visual semantic mapping problem by leveraging advances in foundation models for on-demand, dynamic label generation. Our approach takes posed RGBD data as input and generates panoptic segmentations that form the basis for reconstructing the 3D panoptic volumetric map. In contrast to point-level approaches [20], [21], UPPM, relying on [11], is a superior method for assigning semantics on the objects' level. This makes our dynamically-labeled maps more efficient for map queries fig. 2, which makes UPPM to be used in a wide range of downstream tasks such as localization and navigation. The mapping pipeline is illustrated in fig. 3.

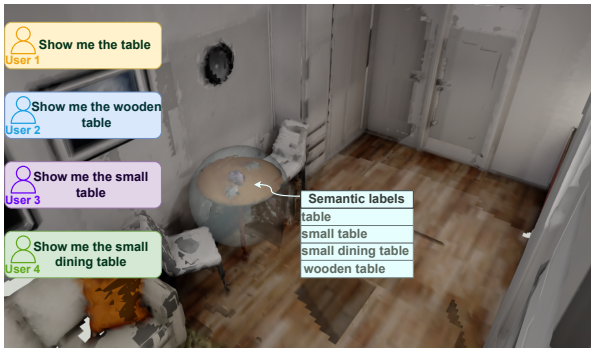


Fig. 2: Our UPPM system enables interactive scene exploration and object retrieval using natural language prompts, employing query postprocessing and STS (Semantic Textual Similarity) for enhanced accuracy. Here, we showcase the ability to employ UPPM in such an application by presenting a scene with a small round wooden table and displaying semantic labels associated with it, showcasing the system's response to four different user prompts, each increasing in specificity.

A. Object-centric 3D Representation

In our study, we build upon the object-centric mapping framework introduced in [11]. Inspired by their semantically consistent representation of the environment, we adopt their notion of dividing the world into submaps for accurate temporal alignment and reduced computational complexity. However, we aim to improve the quality of objects contained within the maps.

The core enhancement in our approach involves refining the basic unit of change: the object. Rather than employing generic entities, we utilize high-quality object instances to generate a richer and more intricate spatial understanding. While preserving the fundamental structure from the original method, we prioritize discerning subtle details inherent in complex environments.

Specifically, every submap encompasses comprehensive information necessary for object tracking, transformations, and label assignments. Panoptic, instance and class labels contribute to complete characterizations of the elements within each submap, facilitating seamless integration with downstream applications requiring precise object recognition and localization capabilities.

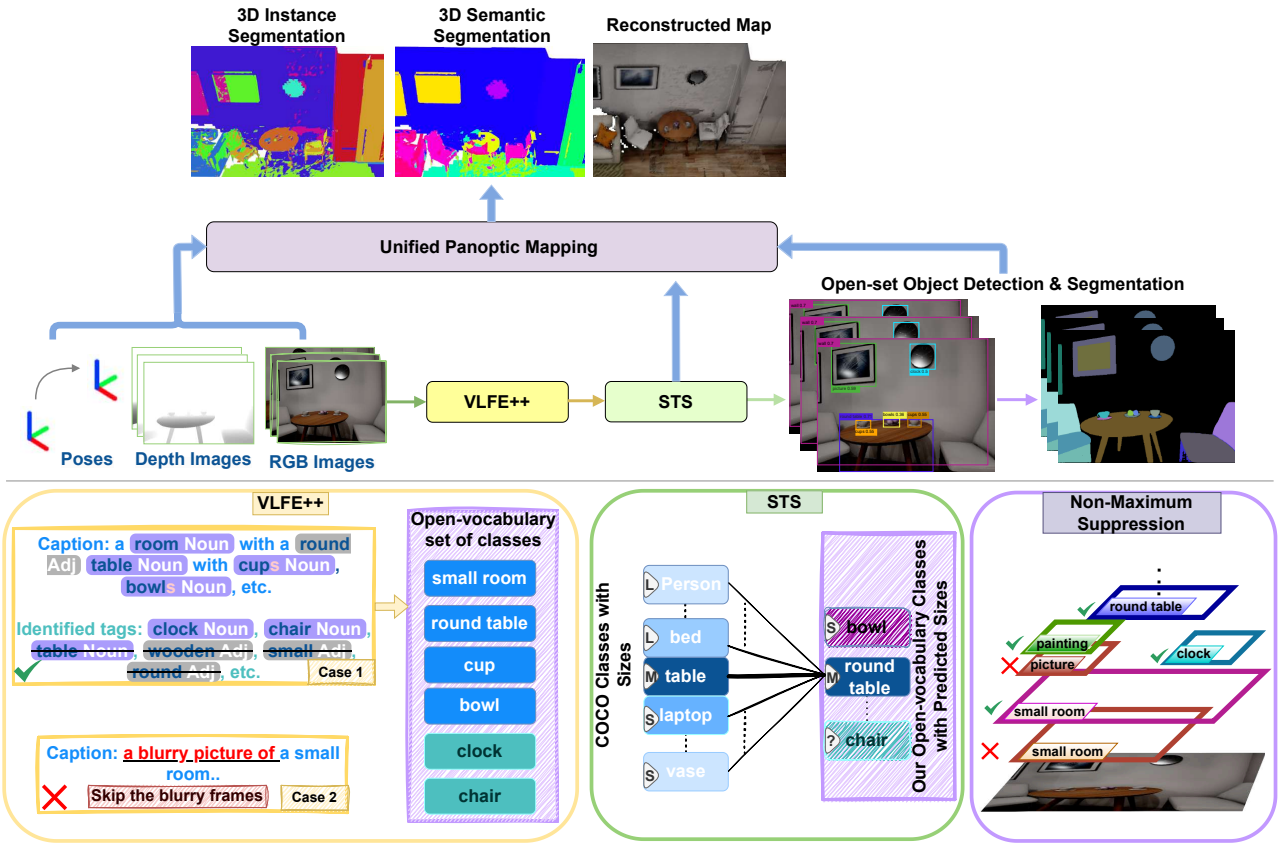


Fig. 3: **System Overview of Unified Promptable Panoptic Mapping (UPPM).** The UPPM pipeline commences with the acquisition of poses, depth images, and RGB images. RGB inputs are processed through VLFE++ (Visual-Linguistic Features Extraction [17], inclusive of part-of-speech tagging [22] and lemmatization [23]), generating detailed object labels enhanced by open-vocabulary prompts for rich semantic content. The Semantic-Textual Similarity (STS) module predicts the size and parent class for each identified object class. This information is then fed into the Open-set Object Detection [19] and Promptable Segmentation module [12], where detection is refined by NMS to ensure heightened accuracy and detail richness. The outputs from the STS module and panoptic segmentation, in conjunction with the posed RGBD inputs, collectively contribute to the reconstruction of the final output. Notably, this reconstruction is influenced by the panoptic mapping approach [11], adapted in our work to accommodate dynamic labels, with a subtle modification ensuring compatibility without overstating the deviation. The final output manifests as a multi-resolution multi-TSDF (Truncated Signed Distance Function) map, fostering natural human-robot communication and enabling versatile applications such as scene exploration and object retrieval using intuitive prompts.

B. Open-Set Classes Generation

In the system overview presented in fig. 3, the VLFE++ comprises various interconnected components designed to generate an open-set classes from an input RGB image. Each component plays a crucial role in the overall process, as described below:

1) **Visual-Linguistic Features Extraction (VLFE):** By harnessing a pre-trained Visual Language Foundation Model such as Tag2Text [17], we encode the input image to derive both visual and semantic information. At this juncture, we obtain significant textual data necessary for subsequent stages.
2) **Part-Of-Speech Tagging (POS tagging):** Applying an average perceptron network [22], we identify potential objects within a scene alongside relevant attributes through grammatical labeling of the generated textual depiction. Here, our focus is particularly on extracting nouns and noun phrases along with any accompanying adjectives, providing rich contextual information about the detected elements present in the given image.
3) **Lemmatization:** Following part-of-speech tagging, lemmatization [23] occurs—a procedure converting inflected or derived word forms into their base or dictionary form (lemma). For instance, instances like ”apples” becoming “apple”, or “chairs” transforming to “chair”. This results in normalized representations of words which subsequently simplifies the extracted terms while preserving their core meanings.

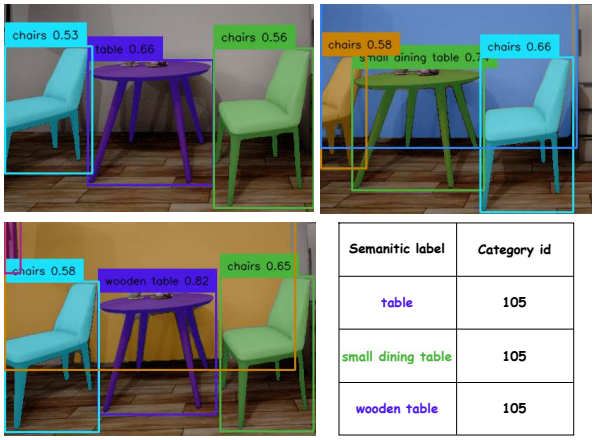
C. Semantic-Textual Similarity (STS)

The task of Semantic-Textual Similarity (STS) aims to quantify the similarity between two pieces of text based on their meaning. In the context of this project, we apply STS fig. 3 to associate size-based categories with corresponding classes in the COCO-Stuff dataset [24]. We define C as the set of all possible classes, where each class $(c_i \in C)$ has a corresponding size attribute (s_i), which takes one of three values - small $(s = 1)$, medium $(s = 2)$, or large $(s = 3)$. Our goal is to determine the most appropriate COCO-Stuff class $\hat{c}$ for any given input class c:

$$\hat{c} = \underset{c' \in C}{\arg\max}\, sim(E(c), E(c')) \tag{1}$$

where $sim(\cdot)$ denotes a similarity function between embeddings and $E(\cdot)$ represents the embedding representation of a class. To generate these embeddings, we follow the method proposed by Song et al. [25]. After obtaining the embeddings, we conduct a semantic search using the algorithm presented by Johnson et al. [26] to select the closest match among the COCO-Stuff classes.

Additionally, we build upon the Panoptic Mapping framework [11] to create a Unified Panoptic Mapping. By incorporating posed RGBD data along with the idea of Unified Semantics, we ensure consistent identification of identical objects across different semantic labels through the assignment of unified category identifiers (fig. 4). This enhancement simplifies classification, leading to a more intuitive interaction experience with spatial information.



Semantic label	Category id
table	105
small dining table	105
wooden table	105

Fig. 4: Unified Semantics in Action: Three sequential frames featuring the same table with distinct semantic labels – "table", "small dining table", and "wooden table". Despite varied descriptions, the unified semantics ensures a consistent category ID across all instances, ensuring semantic cohesion.

D. Open-set Object Detection and Segmentation

Object detection and image segmentation are critical tasks in computer vision, particularly when dealing with unconstrained environments. Our proposed method performs these tasks using curated textual labels as queries for the object detector, thereby facilitating semantic information extraction and object detection. This process results in generated bounding boxes, which then trigger the segmentation model. We employ a promptable zero-shot segmentation model [12] to generate panoptic segmentation maps containing both semantic and instance labels for every pixel within the provided images.

To further refine our approach, we employ Non-Maximum Suppression. During this stage, we address potential redundancies or errors introduced during object detection. Specifically, we focus on rectifying instances involving overlapping bounding boxes assigned identical labels or similar bounding box sizes but dissimilar labels. This approach differs from per-class non-maximum suppression (NMS), which is commonly used in object detection [27]. In mathematical terms:

1) Let B_i denote the set of all bounding boxes extracted from the object detector, such that $|B_i| \geqslant 1$. For any pair (b_a, b_b), where $b_a, b_b \in B_i$, if their Intersection-over-Union (IoU) exceeds a predefined threshold $IoU(b_a, b_b) \geqslant \tau$, and they share the exact class label $y(b_a) = y(b_b)$, one of the duplicates will be suppressed based on criteria like confidence scores.
2) When encountering bounding boxes having equal areas but distinct class labels, the algorithm assigns precedence to the labels originating from captions rather than those derived from previously identified tags. As a result, only one bounding box survives while others get removed through suppression. Mathematically, given two bounding boxes b_c and b_d sharing equivalent dimensions but disparate labels $y(b_c) \neq y(b_d)$, we prioritize retaining b_c whenever its associated label stems from the caption. Otherwise, b_d takes precedence.

E. Implementation Details

We utilized Tag2Text [17] for caption and tag extraction, which provided us with initial descriptions of the environment. To further refine our understanding of open-vocabulary objects and their attributes, we applied part-of-speech (POS) tagging [22] and lemmatization [23]. This allowed us to extract nouns and adjectives from the captions and the identified tags, providing a more precise set of objects and descriptive features.

To embed these objects into a common feature space, we employ MPNet [25] to generate embedding vectors for all relevant COCO-Stuff [24] classes, as well as any open-vocabulary objects identified through our text analysis. Specifically, given an input sequence x containing both open-vocabulary objects & COCO-Stuff classes, MPNet generates corresponding embedding vectors $\mathbf{h} = \text{MPNet}(x)$. This allows us to search for similarities between open-vocabulary objects & COCO-Stuff classes using cosine similarity:

$$sim(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||\ ||\mathbf{v}||}. \tag{2}$$

For each open-vocabulary object, we defined the most similar COCO-Stuff class as its parent class, thereby achieving a unified semantic representation. Moreover, we used the assigned COCO-Stuff class to infer the size estimation for the open-vocabulary object.

Once object detections were generated, we leveraged Grounding-DINO [19] with NMS to ensure that only unique instances were considered. Additionally, we employed Segment Anything model (SAM) [12] to create high-quality instance masks for each detected object. These components with the Panoptic Multi-TSDFs [11] formed the basis of our proposed prompt-based, panoptic mapping pipeline. Rather than introducing a new panoptic segmentation model, our focus was on establishing a dynamic labeling system capable of enriching environmental data with detailed, contextually appropriate metadata.

IV. Experiments

A. Experimental Setup

To evaluate the efficiency of our system, we conducted experiments on both simulated and real-world datasets. The

Flat dataset [11] was utilized for the simulated environment, while evaluations on real-world data were performed using ScanNet v2 [28] and RIO [29] datasets.

For our experiments, a tracking time τ_{new} of 1 frame was assumed for the new submap [11] to be added, ensuring minimal corruption to previously reconstructed data and thereby improving both qualitative and quantitative results.

We measure the performance using root-mean-square error (RMSE), mean-absolute error (MAE), Chamfer distance, and coverage. As shown in eq. (3), the Chamfer distance $d_{\text{ch}}(G, R)$ is computed by combining the sum of RMSE in both directions, $\text{RMSE}_{G\to R}$ and $\text{RMSE}_{R\to G}$, providing the errors from the ground truth point cloud G to the evaluated point cloud R and vice versa.

$$d_{\text{ch}}(G,R) = \underbrace{\sum_{g\in G} \min_{r\in R} \|g-r\|_2^2}_{\text{RMSE}_{G\to R}} + \underbrace{\sum_{r\in R} \min_{g\in G} \|g-r\|_2^2}_{\text{RMSE}_{R\to G}}. \quad (3)$$

Root-mean-square error (RMSE) and mean-absolute error (MAE) are asymmetrical metrics, as the distance from ground truth points (G) to reconstructed map points (R) may differ from R to G. These metrics compute distances by comparing each point in one set to its nearest neighbor in the other. A larger $\text{RMSE}_{R\to G}$ suggests potential inaccuracies in the reconstruction process, emphasizing the importance of accuracy for both sets.

As shown in eq. (4), the coverage is calculated as the percentage of observed ground truth points in the map:

$$\text{Cov} = \frac{N_{\text{observed}}}{N_{\text{total}}} \times 100\%. \quad (4)$$

The experiments were conducted on a system with a 2.6 GHz CPU, an NVIDIA GeForce RTX 2060 6 GB GPU, and 16 GB of RAM, supplemented by a server featuring an NVIDIA GeForce RTX 2080 Ti 16 GB GPU and 32 GB of RAM.

B. Evaluations on the Flat dataset

We evaluate our proposed UPPM model using the Flat dataset [11] as a simulated benchmark, showcasing its effectiveness in scene understanding and reconstruction. Despite its 31-category ground truth segmentation, the Flat dataset provides comprehensive information, highlighting the challenges in achieving full scene understanding, especially for objects like refrigerators, cabinets, and stoves.

We chose this dataset deliberately to highlight the importance of detailed labels for improved scene reconstruction. While the ground truth includes a predefined set of human labels, it might not fully encapsulate the diversity and complexity of real-world scenes. On the other hand, MaskDINO recognizes LVIS1 [30] categories (1203 categories). UPPM, as shown in fig. 3, operates within a customized category list for each new map.

In table I, we present comparative quantitative results demonstrating the superior accuracy and competitive coverage of our UPPM model in comparison to both the ground truth segmentation and the MaskDINO method [4]. Our approach, operating with a customized category list for each new map, emphasizes the significance of incorporating nuanced and diverse labels for robust scene understanding and reconstruction.

Method	GT→Reconst. MAE [m](↓)	GT→Reconst. RMSE [m](↓)	Reconst.→GT MAE [m](↓)	Reconst.→GT RMSE [m](↓)	Chamfer dist. [m](↓)	Cover -age [%](↑)
Groundtruth Seg. [11]	1.27	2.24	0.66	0.80	3.05	71.30
MaskDINO [4]	1.26	2.23	0.68	0.85	3.08	**71.69**
UPPM (Ours)	**1.249**	**2.015**	**0.644**	**0.7854**	**2.8004**	70.85

TABLE I: Quantitative results on the simulated Flat dataset [11]. UPPM exhibits superior accuracy and competitive coverage compared to ground-truth segmentation and MaskDINO [4].

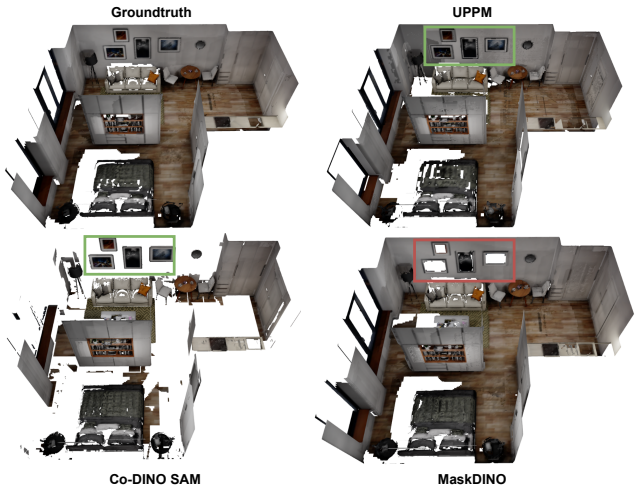


Fig. 5: Qualitative Comparison of Reconstruction Maps on Flat dataset [11].

It is worth noting that while MaskDINO achieves better coverage, a comprehensive comparison between the methods requires examining fig. 5. Upon inspection, it becomes apparent that MaskDINO looks to have better reconstruction for structures like walls and floors. However, it falls short in capturing intricate details present in the scene, like the paintings on the wall, which might hold more semantic value than the wall itself. Conversely, UPPM successfully reconstructs these fine-grained elements without any issues.

C. Evaluations on ScanNet dataset

In order to evaluate the performance of our proposed method, UPPM, we conducted a comparison against MaskDINO [4] and the ground truth segmentation provided by the ScanNet v2 dataset [28]. The ScanNet v2 dataset offers an increased number of categories and enhanced image quality relative to prior datasets, thereby providing a more rigorous evaluation benchmark.

As demonstrated in table II, UPPM exhibits superior accuracy than MaskDINO while maintaining the best coverage in comparison with the ground truth and MaskDINO. However, despite its strong performance, UPPM still falls

short of achieving better accuracy scores than the ground truth segmentation. This discrepancy can be attributed to the inherent complexities present within the ScanNet v2 dataset.

Method	GT→Reconst.		Reconst.→GT		Chamfer dist. [m](↓)	Cover -age [%](↑)
	MAE [m](↓)	RMSE [m](↓)	MAE [m](↓)	RMSE [m](↓)		
Groundtruth Seg. [11]	**1.38**	**2.06**	**2.80**	**7.41**	**9.46**	82.5
MaskDINO [4]	1.81	2.51	4.40	14.26	16.76	81.21
UPPM (Ours)	1.74	2.413	3.656	11.213	13.626	**82.64**

TABLE II: Comparative Evaluation Results on ScanNet v2 Dataset [28].

D. Evaluations on the RIO dataset

Our experiments are carried out using the RIO dataset [29], evaluating the performance of our UPPM model against robust benchmarks like MaskDINO [4] and ground truth segmentation [29]. The RIO dataset presents noisy conditions, creating significant hurdles for our pipeline—particularly regarding limiting error propagation across various pipeline components. Each component is designed to cater to distinct modalities and downstream tasks (section III).

Contrasting our final UPPM implementation with the early Vanilla UPPM approach (without blurry image filtering or NMS) reveals substantial performance gains on the RIO dataset. Our optimized UPPM demonstrates a 16.675% decrease in Chamfer distance and an impressive 6.47% enhancement in coverage. These advancements originate from targeted modifications aimed primarily at enhancing the vanilla UPPM pipeline's ability to handle the noisy characteristics inherent in the RIO dataset.

One major enhancement comes from incorporating an image filtering scheme, removing approximately 3% of RIO dataset images flagged as blurry through caption analysis. As depicted in fig. 3, this approach capitalizes on Tag2Text [17]'s ability to identify blurred images. By eliminating low-quality imagery, our method brings UPPM outputs closer to the ground truth segmentations, fig. 6.

Fig. 6: **Comparative Visualization. Top Row:** Groundtruth segmentation showcasing the spatial distribution (Right) and its corresponding reconstructed colored map (Left) for enhanced scene comprehension. **Bottom Row:** Our proposed UPPM approach, demonstrating refined segmentation with pronounced clarity and detail in both the spatial (Right) and colored map (Left) representations, underscoring the superior performance of UPPM in intricate scene understanding.

To better understand our contributions, we compare the final UPPM to MaskDINO [4] and ground truth segmentation [29] on the RIO dataset [29], summarizing our observations in table III. Unlike the Flat and ScanNet datasets, MaskDINO [4] seems to have better performance on the RIO dataset, which we attribute to the MaskDINO closed-set classes seems to work well on RIO dataset which might be not the case in other unseen environments.

Method	GT→Reconst.		Reconst.→GT		Chamfer dist. [m](↓)	Cover -age [%](↑)
	MAE [m](↓)	RMSE [m](↓)	MAE [m](↓)	RMSE [m](↓)		
Groundtruth Seg. [11]	**1.23**	**1.68**	1.91	8.57	10.25	**77.09**
MaskDINO [4]	1.28	1.74	1.97	**7.42**	**9.16**	76.63
UPPM (Ours)	1.315	1.772	**1.87**	8.052	9.824	75.14

TABLE III: Performance comparison of our final UPPM model, MaskDINO, and ground truth segmentation on the RIO dataset.

E. Ablation Studies

We conduct ablation studies on each dataset independently to gain a deeper understanding of the contributions made by different components in our UPPM model. The objective of these studies is to offer insights into the trade-offs between accuracy and generalizability across diverse scenarios by examining the impacts of individual components and their variations within the UPPM architecture. We classify our ablation experiments into four main categories:

Firstly, we investigate the performance of the **UPPM w/o tags**. In this setup, we exclude additional tags from the UPPM model (fig. 3). Despite some debate around the utility of these tags, as they may not guarantee the presence of related objects in captions, our findings indicate that such tags typically correspond to real objects existing in Flat dataset images. As a result, we regard them as beneficial supplementary inputs.

Secondly, we study the effect of disabling the unified semantic mechanism (**PPM**) while retaining the extra tags. By doing so, we intend to determine if the unified semantics have any detrimental influence on map reconstruction, whilst still benefiting from the added tags.

Thirdly, in **PPM w/o tags**, we deactivate both the unified semantics mechanism and the extra tags, allowing us to explore possible negative consequences associated with these features during map reconstructions.

Lastly, we replace the open-set object detector with a closed-set counterpart, specifically, Co-DINO [5], referred to as **Co-DINO+SAM**. With Co-DINO identifying all LVIS1 [30] classes comprised of 1203 unique categories, we analyze how it affects the overall performance of the system.

Below, you find the results and detailed analyses of the above experiments conducted on three datasets:

*1) **Flat dataset ablation experiments**:* As shown in table IV, our ablation studies show that both UPPM and PPM perform similarly on the Flat dataset. However, Co-DINO+SAM excels with lower errors in Chamfer distance

and all other metrics, excluding coverage. Although Co-DINO+SAM yields impressive quantitative results, it faces challenges when precisely identifying background objects during recognition tasks (see fig. 5). Notably, our UPPM approach competes well against the other settings by giving the combination of dynamic-labeling, competitive accuracy and high coverage.

Method	GT→Reconst. MAE [m](↓)	GT→Reconst. RMSE [m](↓)	Reconst.→GT MAE [m](↓)	Reconst.→GT RMSE [m](↓)	Chamfer dist. [m](↓)	Cover -age [%](↑)
Co-DINO+SAM	0.90	1.39	0.62	0.73	2.12	44.84
PPM w/o tags	1.207	1.994	0.651	0.8037	2.7977	70.38
PPM	1.233	2.017	0.643	0.8032	2.8202	71.26
UPPM w/o tags	1.25	2.065	0.652	0.8103	2.8753	70.76
UPPM	1.249	2.015	0.644	0.7854	2.8004	70.85

TABLE IV: Quantitative Ablation Experiments on the Flat dataset [11].

*2) **Ablation Experiments on ScanNet Dataset**:* In this section, we discuss the findings from our ablation studies conducted using the ScanNet v2 dataset [28]. The proposed synthetic approach, Co-DINO+SAM, demonstrates the best performance concerning the Chamfer distance metric. However, there remains room for improvement regarding coverage when compared to alternative techniques. Upon further investigation, we attribute this disparity to limitations arising from the closed-set design of the model and restrictions related to training data.

Although PPM and UPPM exhibit slightly reduced accuracy due to imperfect image quality inherent within the ScanNet dataset (0.84% of images were flagged as blurry, significantly less than those found in RIO), both metrics yield comparable and remarkably high levels of coverage. Moreover, their corresponding Chamfer distances remain relatively close to the ground truth segmentation (table II), signifying robustness against real-world challenges posed by datasets such as ScanNet.

Method	GT→Reconst. MAE [m](↓)	GT→Reconst. RMSE [m](↓)	Reconst.→GT MAE [m](↓)	Reconst.→GT RMSE [m](↓)	Chamfer dist. [m](↓)	Cover -age [%](↑)
Co-DINO+SAM	1.51	2.03	2.25	5.43	7.46	59.3
PPM w/o tags	1.794	2.477	3.064	9.166	11.643	81.07
PPM	1.736	2.395	3.365	10.106	12.501	**82.75**
UPPM w/o tags	1.781	2.432	3.299	10.21	12.642	81.42
UPPM	1.74	2.413	3.656	11.213	13.626	**82.64**

TABLE V: Quantitative Ablation Experiments on the ScanNet v2 [28].

One might wonder why UPPM should be used when PPM offers higher coverage and superior accuracy. However, as depicted in fig. 7, the reason becomes apparent. While PPM boasts an open-vocabulary, it lacks dynamically labeled classes, resulting in potential ambiguity where multiple semantic categories may apply to the same object. This can result in reduced control over the environment due to duplicated classes assigned to the same object. In contrast, UPPM demonstrates a more consistent behavior by accurately assigning semantic classes to objects while maintaining richness in dynamic labeling.

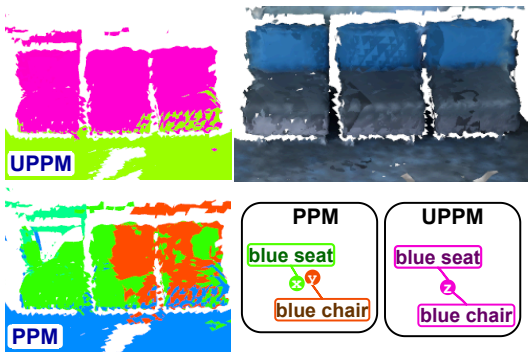


Fig. 7: Qualitative Comparison of PPM vs UPPM in 3D-Semantic Segmentation. The predicted semantic classes are represented by $x, y, \& z$.

*3) **RIO dataset ablation experiments**:* The ablation experiments, as shown in table VI, provide valuable insights on on how different factors may affect the algorithm performance. Specifically, we observed that both UPPM w/o tags ad UPPM w/o tags demonstrated the best coverage across all experiments while maintaining competent accuracy. This result highlights the negative effects of poor image quality and motion blur on the performance of UPPM with tags, indicating that these issues may lead to suboptimal results.

Interestingly, we notice that the Tag2Text algorithm [17] identifies approximately 3% of the RIO dataset as containing blurry images. However, manual inspection and classification of the data revealed that over 21% of the dataset is affected by motion blur or other forms of degradation. This discrepancy suggests that Tag2Text identifies only extreme cases of blurriness, leaving many less severe cases undetected.

On the other hand, the Co-DINO+SAM model exhibited the largest Chamfer distance among all models evaluated. We attribute this behavior to substantial deviations from the ground truth during map reconstruction, leading to reduced accuracy and reliability.

Based on the findings obtained through our analysis of the RIO dataset, we draw the following conclusion: In scenarios where the dataset contains noise or poor quality data, relying solely on the UPPM w/o tags seems to yield better performance. By doing so, one maintains the benefits of dynamic labeling and unified semantics while preventing low-quality tags from contaminating the input fed to the open-set object detector [19].

Method	GT→Reconst. MAE [m](↓)	GT→Reconst. RMSE [m](↓)	Reconst.→GT MAE [m](↓)	Reconst.→GT RMSE [m](↓)	Chamfer dist. [m](↓)	Cover -age [%](↑)
Co-DINO+SAM	1.11	1.46	2.76	14.55	16.01	43.42
PPM w/o tags	1.2339	1.699	1.38	2.559	4.258	72.47
PPM	1.2332	1.698	1.73	7.304	9.002	73.49
UPPM w/o tags	1.195	1.628	1.547	4.024	5.652	74.14
UPPM	1.315	1.772	1.87	8.052	9.824	**75.14**

TABLE VI: Quantitative Ablation Experiments on the RIO data [29].

V. Conclusion

In this work, we propose UPPM, a novel approach that tackles the challenges of generating rich and accurate object labels for panoptic mapping by harnessing the power of dynamic labeling. Our system efficiently integrates diverse, possibly noisy labels from multiple sources into a consistent semantic structure, leading to effective postprocessing and precise 2D segmentation. We demonstrated the usefulness of foundation models and their efficiency in being utilized in downstream mapping tasks without requiring model retraining. Furthermore, our findings highlight the limitations of model outputs and emphasize the importance of postprocessing model outputs to ensure optimal performance in real-world scenarios.

References

[1] G. Narita, T. Seno, T. Ishikawa, and Y. Kaji, "Panopticfusion: Online volumetric semantic mapping at the level of stuff and things," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4205–4212, IEEE, 2019.

[2] J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger, "Fusion++: Volumetric object-level slam," in *2018 international conference on 3D vision (3DV)*, pp. 32–41, IEEE, 2018.

[3] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1352–1359, 2013.

[4] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3041–3050, 2023.

[5] Z. Zong, G. Song, and Y. Liu, "Detrs with collaborative hybrid assignments training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6748–6758, 2023.

[6] K. Mazur, E. Sucar, and A. J. Davison, "Feature-realistic neural fusion for real-time, open set scene understanding," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8201–8207, IEEE, 2023.

[7] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, A. Maalouf, S. Li, G. Iyer, S. Saryazdi, N. Keetha, *et al.*, "Conceptfusion: Open-set multimodal 3d mapping," *arXiv preprint arXiv:2302.07241*, 2023.

[8] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *2017 IEEE International Conference on Robotics and automation (ICRA)*, pp. 4628–4635, IEEE, 2017.

[9] Y. Xiang and D. Fox, "Da-rnn: Semantic mapping with data associated recurrent neural networks," *arXiv preprint arXiv:1703.03098*, 2017.

[10] N. Merrill, Y. Guo, X. Zuo, X. Huang, S. Leutenegger, X. Peng, L. Ren, and G. Huang, "Symmetry and uncertainty-aware object slam for 6dof object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14901–14910, 2022.

[11] L. Schmid, J. Delmerico, J. L. Schönberger, J. Nieto, M. Pollefeys, R. Siegwart, and C. Cadena, "Panoptic multi-tsdfs: a flexible representation for online multi-resolution volumetric mapping and long-term dynamic scene consistency," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 8018–8024, IEEE, 2022.

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[14] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*, pp. 8821–8831, PMLR, 2021.

[15] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

[16] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[17] X. Huang, Y. Zhang, J. Ma, W. Tian, R. Feng, Y. Zhang, Y. Li, Y. Guo, and L. Zhang, "Tag2text: Guiding vision-language model via image tagging," *arXiv preprint arXiv:2303.05657*, 2023.

[18] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *European Conference on Computer Vision*, pp. 350–368, Springer, 2022.

[19] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[20] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[21] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "imap: Implicit mapping and positioning in real-time," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6229–6238, 2021.

[22] M. Honnibal, "A good part-of-speech tagger in about 200 lines of python." https://explosion.ai/blog/part-of-speech-pos-tagger-in-python, 2013.

[23] C. Fellbaum, "Wordnet," in *Theory and applications of ontology: computer applications*, pp. 231–243, Springer, 2010.

[24] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

[25] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16857–16867, 2020.

[26] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[28] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

[29] J. Wald, A. Avetisyan, N. Navab, F. Tombari, and M. Nießner, "Rio: 3d object instance re-localization in changing indoor environments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7658–7667, 2019.

[30] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364, 2019.