A Category-Specific Prompt Strategy for Semantic 3D Indoor Mapping Using RGB-D Camera

Jiwei Hou[1], Vivien Volland[2], Samer Karam[1], Dorota Iwaszczuk[1]

[1] Remote Sensing and Image Analysis, Department of Civil and Environmental Engineering, Technical University of Darmstadt, 64287 Darmstadt, Germany - (jiwei.hou, samer.karam, dorota.iwaszczuk)@tu-darmstadt.de
[2] Geodetic Measurement Systems and Sensor Technology, Department of Civil and Environmental Engineering, Technical University of Darmstadt, 64287 Darmstadt, Germany - vivien.volland@tu-darmstadt.de

Keywords: 3D Semantic Mapping, RGB-D SLAM, Segment Anything Model, Prompt-Based Segmentation, Indoor Mapping.

Abstract

Semantic 3D indoor mapping often depends on supervised learning and large annotated datasets, limiting scalability across diverse environments. This work introduces a category-specific prompt strategy for semantic 3D mapping using RGB-D cameras, integrating RGB-D SLAM with the Segment Anything Model 2 (SAM2) to enable annotation-efficient reconstruction. Keyframes and trajectories extracted from SLAM provide spatial references, while SAM2 performs zero-shot segmentation guided by a Category-Wise Prompt Segmentation Strategy (CPSS), which segments structural and functional elements (e.g., floors, doors, staircases) by category to reduce prompt interference and manual effort. The segmented keyframes are then fused with depth and pose data to produce instance-level semantic point clouds. Experiments on custom RGB-D sequences and selected ScanNet scenes demonstrate centimeter-scale geometric consistency and strong semantic consistency, with mIoU values up to 0.89 on the custom dataset and 0.98 on ScanNet. The resulting semantic point clouds are clean, structured, and require minimal post-processing, showing that the proposed strategy provides an efficient and scalable solution for semantic 3D indoor mapping without retraining or environment-specific supervision.

1. Introduction

Constructing accurate and semantically rich 3D maps is essential for many indoor applications, including digital twins, building modeling, and facility management. Such maps integrate geometric structure with semantic labels, enabling downstream tasks such as building information modeling (BIM), spatial analysis, and robotic operation in general. With the widespread use of low-cost RGB-D sensors and recent advances in prompt-based segmentation, it is now feasible to explore scalable strategies for producing semantic maps without large annotated datasets.

However, generating such maps in a scalable and robust manner remains challenging. Existing 3D semantic mapping methods primarily rely on supervised deep learning models that require large annotated datasets (Qi et al., 2017a, Qi et al., 2017b), which are costly to produce, especially in large or dynamic indoor environments. Although unsupervised (Landrieu and Simonovsky, 2018), self-supervised (Xie et al., 2020), and semi-supervised methods (Xu and Lee, 2020) reduce annotation needs, they often struggle to achieve fine-grained accuracy or maintain robust generalization across diverse scenes.

Prompt engineering has recently emerged as a promising way to guide large-scale models toward specific tasks without retraining. In computer vision, the Segment Anything Model (SAM) (Kirillov et al., 2023, He et al., 2025) and its improved version SAM2 (Ravi et al., 2024) enable zero-shot segmentation based on simple prompts (e.g., points, boxes, or text) and achieve strong generalization even in unseen scenes. This opens new possibilities for low-cost segmentation of indoor elements. Early attempts to extend the original SAM to 3D scenes include SAM3D (Yang et al., 2023), which aggregates SAM-based 2D masks across multiple views. However, current prompt-based segmentation operates mainly on 2D images. Projecting 2D results into 3D requires accurate depth alignment across frames, which introduces high computational overhead and error accumulation when processing all RGB-D frames.

Therefore, selecting keyframes with reliable camera poses is crucial to reduce computation while maintaining accuracy in continuous RGB-D sequences (Yarovoi and Cho, 2024). In this work, we leverage RGB-D SLAM frameworks to obtain accurate camera trajectories and keyframes for efficient 3D semantic mapping. Specifically, ORB-SLAM3 (Campos et al., 2021) is used for our custom dataset, while the ScanNet benchmark provides precomputed poses from BundleFusion (Dai et al., 2017b). While semantic maps can support many tasks—including navigation, BIM conversion, and digital twin construction—in this work we focus on the core methodological problem: how to efficiently construct instance-level 3D semantic maps from RGB-D video with minimal annotation and no retraining.

The main contribution of this work is a prompt-based strategy for 3D semantic mapping of indoor structural and functional elements using RGB-D video data. The proposed framework leverages zero-shot segmentation and RGB-D SLAM–derived poses to generate accurate instance-level semantic point clouds with minimal manual prompting and no task-specific retraining. To further improve segmentation consistency, we introduce a Category-Wise Prompt Segmentation Strategy (CPSS), which performs prompt-based segmentation separately for each semantic category rather than jointly across all classes. Experimental results demonstrate that the proposed strategy effectively mitigates cross-category interference, reduces prompt effort, and improves the overall reliability of 3D semantic labeling. Furthermore, CPSS offers a practical and scalable workflow for applying prompt-based segmentation to large-scale indoor semantic mapping.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-255-2026

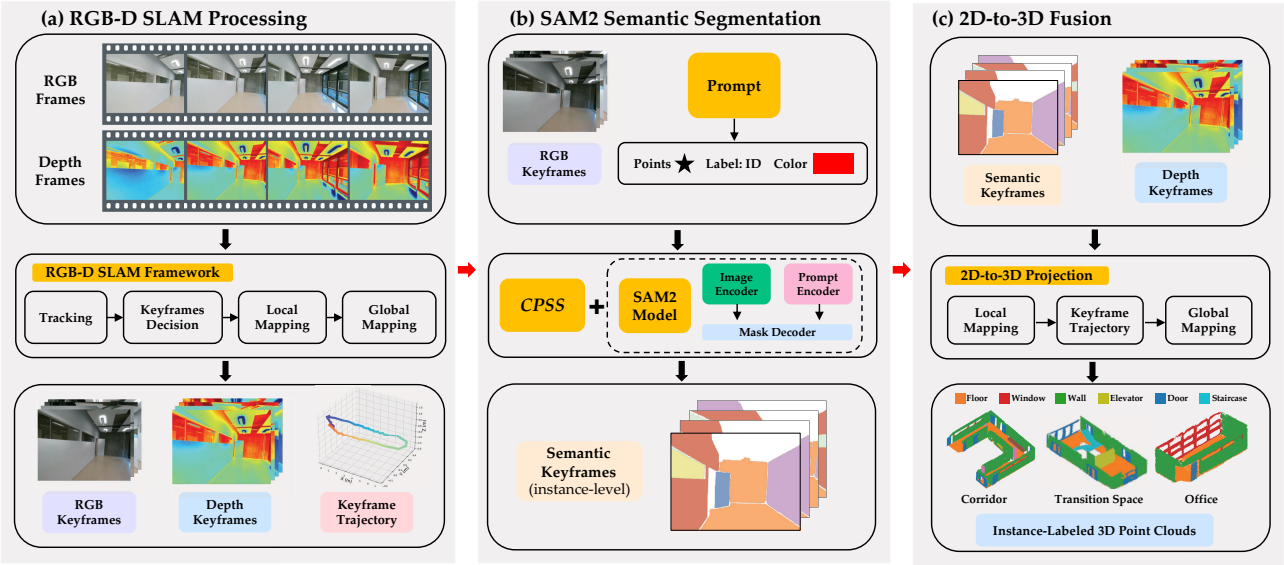

Figure 1. Overview of the proposed pipeline: (a) keyframe and trajectory extraction using RGB-D SLAM; (b) prompt-based semantic segmentation of RGB keyframes with SAM2 using the CPSS strategy; (c) 2D–3D fusion to generate instance-labeled 3D point clouds.

2. Related Work

Semantic 3D indoor mapping has been approached by extending SLAM pipelines with semantic perception. In these methods, object detectors or segmentation networks are integrated into RGB-D SLAM (Campos et al., 2021) to enrich geometric maps with semantic labels (Bescos et al., 2018, Chen et al., 2023). While effective, such systems rely on supervised models requiring large annotated datasets and often generalize poorly to unseen indoor environments. Furthermore, coupling semantic inference with tracking increases computational overhead and limits scalability in dense or long-term mapping scenarios.

An alternative direction is to perform semantic segmentation directly on 3D point clouds. Deep-learning-based approaches such as PointNet++ learn hierarchical geometric features (Qi et al., 2017b), while unsupervised and weakly supervised variants (Xie et al., 2020) aim to reduce annotation requirements. However, dense 3D segmentation remains computationally demanding and difficult to scale to large indoor spaces. In addition, existing models often struggle with fine-grained instance differentiation and semantic consistency, especially for indoor structural and functional elements that exhibit similar geometry or appear in close proximity.

In recent years, prompt engineering has emerged as a promising way to reduce annotation dependency by guiding large-scale pre-trained models toward specific tasks without retraining (Wang et al., 2023, Brown et al., 2020). In computer vision, the Segment Anything Model (SAM) (Kirillov et al., 2023) enables zero-shot segmentation from simple prompts such as points, boxes, or text. Its successor SAM2 (Ravi et al., 2024) extends this capability to video with improved temporal consistency and segmentation quality, making it suitable for RGB-D sequences in indoor environments.

Recent studies have explored integrating SAM-based 2D segmentation with 3D scene reconstruction. For example, SAI3D (Yin et al., 2024) combines SAM-generated masks with geometric priors for zero-shot 3D segmentation, while other works project 3D point clouds into 2D views to leverage the strong generalization capability of SAM (Kang et al., 2024, Wang et al., 2022). SAM2Point (Guo et al., 2024) further converts point clouds into voxelized video representations for SAM2-based segmentation. However, these approaches remain constrained by view dependency, geometric detail loss during projection, and high computational overhead, which limit their scalability for complex indoor mapping tasks.

To address these limitations, we propose an offline approach for indoor 3D semantic mapping that emphasizes scalability and minimizes manual effort. By applying category-specific prompts to structural and functional elements (e.g., doors, walls, staircases), the method derives reliable instance-level semantics from RGB video sequences without environment-specific retraining.

3. Methodology

The workflow of the proposed method is illustrated in Figure 1, with each step detailed in Sections 3.1–3.3.

3.1 RGB-D SLAM Processing

This stage uses an existing RGB-D SLAM module to obtain camera trajectories and keyframes as spatial references for subsequent semantic processing. For the custom RGB-D dataset, ORB-SLAM3 (Campos et al., 2021) is adopted as the localization and mapping backbone. An extended version of ORB-SLAM3 (Hou et al., 2023), which supports dense point-cloud reconstruction and is adapted to the Intel RealSense D455 sensor, is employed to process the RGB-D streams without modifying the core algorithm. The resulting keyframes and camera poses provide the spatial foundation for subsequent semantic segmentation and 3D fusion.

For experiments on the ScanNet dataset (Dai et al., 2017a), we do not re-run SLAM. Instead, we use the precomputed camera poses provided by BundleFusion (Dai et al., 2017b). To reduce redundancy in 3D reconstruction, keyframes are extracted from ScanNet sequences by referencing the keyframe selection principles of ORB-SLAM3. Specifically, a new keyframe is added

when the camera translates more than 0.10 m, rotates over 10°, or after 20 frames, while frames with a valid depth ratio below 0.3 within the 0.2–4.0 m range are discarded.

In summary, the SLAM module is an off-the-shelf component that provides reliable spatial and temporal alignment, while the subsequent stages—prompt-based segmentation and 2D-to-3D fusion—constitute the main methodological contributions of this work.

3.2 SAM2-based Semantic Segmentation

After obtaining RGB keyframes, depth keyframes, and camera trajectories based on the method presented in Section 3.1, the next stage leverages RGB keyframes as input to SAM2 for prompt-based interactive segmentation. In this work, SAM2 is used in its original form, and our contribution focuses on the category-wise organization of prompts. Here, points are used as prompts, as supported by SAM2 (Ravi et al., 2024), which performs zero-shot segmentation from simple user inputs (e.g., points, boxes, or text) and generalizes well even in unseen scenes.

However, prompting multiple heterogeneous categories within the same frame increases prediction and tracking uncertainty, which may cause ambiguous boundaries, segmentation drift, or incorrect mask propagation.

To achieve more accurate instance-level segmentation of indoor structural and functional elements, we propose a Category-Wise Prompt Segmentation Strategy (CPSS), where different element categories are segmented independently, for example, only doors are segmented when extracting doors, and only walls are considered when extracting walls. The hierarchical relationships among these categories and their corresponding instances are illustrated in Figure 2. Finally, temporally consistent instance masks (masklets) from all categories are merged to produce a complete 2D instance-level semantic mask.

Taking doors in a corridor as an example, the segmentation process consists of two key stages, as illustrated in Figure 3. First is the initial selection. In the first keyframe, all visible doors are individually prompted to ensure instance-level segmentation. As shown in Figure 3, green markers are used as positive prompts to specify the target doors, while red markers serve as optional negative prompts to refine the segmentation by excluding adjacent or overlapping structures when finer granularity is required. Each successfully segmented door is assigned a unique instance ID to distinguish it from other objects of the same category. Once all visible doors are accurately identified in the first frame, SAM2 propagates the segmentation results across continuous frames (blue arrows), forming masklets that maintain temporal consistency. However, when segmentation quality degrades due to occlusions, illumination changes, or complex scene geometry, SAM2 may lose track of the object in subsequent frames. In such cases, additional prompts can be manually provided in a later frame (red arrow) to restore accuracy and continue the propagation.

Then, during the refinement stage, if segmentation inconsistencies arise, additional prompts in a later frame (e.g., Frame 3) are added to recover the lost segmentation and resume propagation. When a previously unseen door appears in the scene, a new prompt is added to initialize its segmentation, and a new unique instance ID is assigned to ensure it is tracked as a distinct object. This correction and initialization mechanism significantly reduces annotation effort compared to conventional segmentation methods. By leveraging the memory of SAM2, the proposed method enables efficient and robust instance-level segmentation with minimal manual intervention, ensuring improved accuracy and continuity in large-scale indoor environments.

Although CPSS segments categories sequentially, it reduces the overall interaction effort by avoiding error propagation and repeated corrections that often occur when prompting multiple heterogeneous objects simultaneously. Through this interactive segmentation strategy, we obtain temporally consistent masklets of large-scale structural elements in indoor environments. These segmented keyframes provide clean, well-structured semantic inputs, serving as a reliable foundation for subsequent 2D-to-3D fusion in the semantic indoor mapping process.

An ablation study is conducted to evaluate the contribution of the proposed CPSS strategy. On the ScanNet dataset, we compare multi-category and category-wise prompting in terms of mean IoU, corrections, and prompt effort, aiming to demonstrate the effectiveness of CPSS in improving segmentation consistency and efficiency.

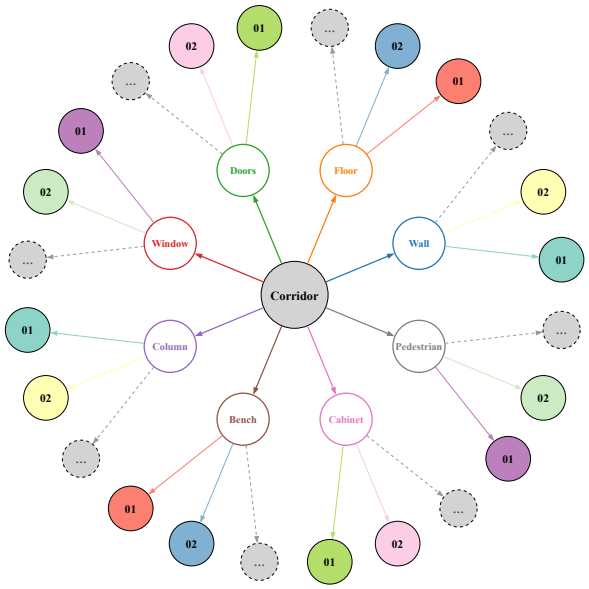

Figure 2. Category-wise organization of indoor elements in a dynamic corridor scene, with categories in the inner layer and instances in the outer layer.

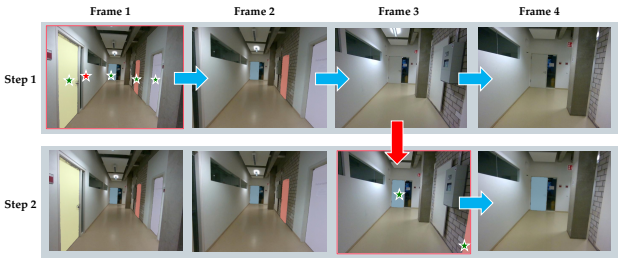

Figure 3. Prompt-based instance-level segmentation with SAM2: a case study on corridor doors.

3.3 2D-to-3D Fusion

The 2D-to-3D fusion module integrates outputs from the previous stages—semantic keyframes, depth keyframes, and key-

frame trajectories—to generate instance-level 3D semantic point clouds of indoor environments.

Each depth keyframe is reprojected into the global coordinate frame using the corresponding camera pose and intrinsic parameters. Invalid depth values are filtered by constraining the valid range between 0.1 m and 5 m to remove out-of-range measurements. For each valid pixel coordinate (u, v) in the RGB keyframe with depth $D(u, v)$, the corresponding 3D point (X, Y, Z) in the camera coordinate system is computed as:

$$X = \frac{(u - u_0)D(u, v)}{f_x}, Y = \frac{(v - v_0)D(u, v)}{f_y}, Z = D(u, v) \tag{1}$$

where (u_0, v_0) are the principal point coordinates and (f_x, f_y) are the focal lengths.

The local point clouds reconstructed from individual keyframes are then aligned into a global coordinate system using their associated keyframe poses. During this process, the semantic masks generated by CPSS (e.g., doors, walls, staircases) are projected into 3D space, assigning both semantic category labels and unique instance IDs to the corresponding 3D points.

Ambiguous or overlapping 2D mask regions are filtered prior to 2D-to-3D projection to avoid label conflicts. After mask merging and filtering, 2D masks are projected into 3D space separately for each semantic category. Voxel grid downsampling with a voxel size of 2 cm is then applied independently to each category-specific point cloud to balance point density and structural detail.

A subsequent lightweight post-processing step removes outliers and residual noise introduced by sensor depth errors. The resulting clean, instance-level semantic point cloud serves as the final output of the mapping framework, enabling downstream tasks such as geometry modeling, structural analysis, and indoor spatial reasoning.

4. Experiments and Results

The experiments were conducted primarily in a Google Colab environment equipped with an NVIDIA A100 GPU (40 GB VRAM) and approximately 90 GB of system RAM. Point cloud post-processing was performed in a standard Ubuntu environment. All ablation experiments were conducted by the same operator to ensure consistent interaction conditions.

4.1 Experimental Dataset

ScanNet (Dai et al., 2017a) is a widely used RGB-D benchmark for evaluating semantic mapping methods, but it focuses mainly on object-level understanding and offers limited coverage of large structural elements such as corridors and staircases. To better target these components, we constructed a custom indoor dataset tailored for structural semantic mapping. For generalizability in public benchmarks, we evaluated our method on selected ScanNet scenes.

For the custom dataset, three representative indoor scenes were recorded: a corridor, a transition space containing staircases and elevator areas, and an office, as illustrated in Figure 4. In addition, a dynamic corridor sequence with moving pedestrians was

Figure 4. Illustration of the customized indoor environments used in our experiments. All scenes were recorded using an Intel RealSense D455 RGB-D camera (Hübner et al., 2023).

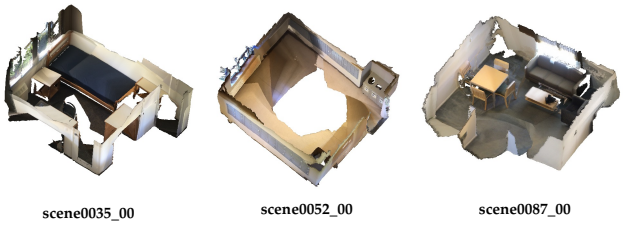

Figure 5. Representative indoor scenes from the ScanNet dataset used for ablation experiments.

captured to simulate real-world occlusions and dynamic interference. Table 1 summarizes the RGB-D recordings used in the experiments.

To complement the custom dataset and assess the generalization capability of the proposed framework, we further evaluated our method on three representative indoor scenes from the ScanNet dataset (Dai et al., 2017a). The dataset provides RGB-D image sequences and globally optimized camera trajectories computed by BundleFusion (Dai et al., 2017b). The selected scenes include a bedroom (*scene0035_00*), an unfurnished room (*scene0052_00*), and a living room (*scene0087_00*), representing diverse spatial layouts and object densities. Figure 5 illustrates the reconstructed meshes of these scenes used for the ablation study between multi-category prompting and the proposed CPSS strategy.

4.2 Evaluation Metrics

We evaluate the reconstructed semantic point clouds from two perspectives. First, relative geometric alignment is assessed by computing the cloud-to-cloud distance between the reconstructed point clouds and reference point clouds acquired with an industrial-grade laser scanning system. The mean distance and standard deviation are reported as alignment metrics.

Second, segmentation quality is evaluated using the Intersection over Union (IoU), computed per semantic class between the reconstructed and manually annotated ground truth point clouds:

$$IoU = \frac{|P \cap G|}{|P \cup G|}, \tag{2}$$

where P and G denote the reconstructed and ground truth point sets, respectively. A reconstructed point $p_i \in P$ is considered overlapping with a ground truth point $g_j \in G$ if their Euclidean

Scene	Area (m²)	Duration (s)	Trajectory (m)	Keyframes	Avg. Prompt Frames / Instance	Avg. Points / Prompt Frame
Static Corridor	67	95	33.85	167	1.32	3.94
Transition Space	107	113	48.45	191	2.42	4.76
Office	58	63	20.44	92	1.00	3.25
Dynamic Corridor	67	94	34.23	168	1.19	4.06

Table 1. Summary of the custom RGB-D recordings used in the experiments.

distance satisfies:

$$\|p_i - g_j\|_2 \leq \tau_{IoU}, \tag{3}$$

where τ_{IoU} is the overlap threshold.

To ensure consistency across all scenes, τ_{IoU} was defined as three times the average standard deviation of the cloud-to-cloud distances computed between the reconstructed and ground-truth point clouds across all scenes, resulting in a unified threshold of $\tau_{IoU} = 0.141\,m$.

It should also be noted that the reconstructed scenes were evaluated only on elements consistently present in both the reconstruction and the ground-truth data. For example, in the transition space, the doors had been modified during subsequent renovations, and their states (e.g., open vs. closed) differed from those in the ground-truth scan. To ensure fair evaluation, door regions in this scene were excluded from the comparison.

For the ScanNet dataset, the same evaluation procedure was applied. Since the provided camera poses from BundleFusion ensure highly accurate spatial alignment, the focus was placed on assessing semantic segmentation performance rather than geometric reconstruction accuracy. The ground-truth annotations provided by ScanNet were used for semantic evaluation, with a unified overlap threshold of $\tau_{IoU} = 0.050\,m$ applied to all scenes.

4.3 Visualization of Results

Figure 6 shows representative SAM2-based 2D segmentation results produced with the proposed CPSS strategy. Instance masklets are overlaid on RGB frames, and sequential examples are shown across different environments. The segmented keyframes produced by CPSS yield clean and temporally consistent semantic inputs for 3D fusion. Figure 7 illustrates examples of the resulting 3D semantic maps and their comparison with ground truth.

Figure 8 presents additional qualitative results on ScanNet, complementing the quantitative analysis in Table 3. The ablation results exhibit confusion between cabinet and table classes in *scene0035_00* and incomplete wall segments in *scene0052_00* and *scene0087_00*. In contrast, CPSS yields cleaner and more coherent structures. Interestingly, in certain regions such as walls, doors, and radiators, both methods even exhibit sharper boundaries than the ScanNet ground-truth annotations.

4.4 Evaluation of the Results

Evaluation on the custom dataset. According to the metrics in Section 4.2, the mean cloud-to-cloud distances for the static corridor, transition space, office, and dynamic corridor are 0.036 m, 0.042 m, 0.104 m, and 0.045 m, with standard deviations of 0.031 m, 0.040 m, 0.081 m, and 0.036 m. These values indicate centimeter-scale geometric consistency with the ground truth across most scenes.

Figure 6. Representative 2D segmentation results using the proposed CPSS strategy. Instance masklets are overlaid on RGB frames, with two-row sequences shown for each scene.

Quantitative and qualitative segmentation results are summarized in Table 2 and Figure 7. The static and dynamic corridors both achieve 0.89 mIoU, showing that dynamic elements have limited impact on segmentation quality. The transition space attains 0.87 mIoU, while the office yields a lower 0.76 mIoU, mainly due to depth-sensing noise on glass surfaces and wall gaps caused by door occlusions.

Figure 7 further illustrates the reconstructed semantic point clouds compared with ground truth, showing high completeness for major structural and functional elements, including floors, walls, doors, staircases, and the elevator.

Evaluation on the ScanNet dataset. To further validate the generalization of the proposed framework, additional experiments were conducted on three representative ScanNet scenes (*scene0035_00*, *scene0052_00*, and *scene0087_00*).

Table 3 compares the prompt effort and correction requirements between multi-category prompting (Ablation) and the proposed CPSS strategy. Correction Frames correspond to the refinement steps described in Section 3.2, where additional prompts were added to recover segmentation drift or lost masks. With substantially fewer prompt frames and points, CPSS achieves comparable or higher IoU scores. On average, CPSS required 1.2 prompt frames per instance and 4.7 prompt points per prompt frame, while the ablation setting required 1.3 prompt frames

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-255-2026 | © Author(s) 2026. CC BY 4.0 License.

Scene	Floor	Door	Wall	Window	Staircase	Elevator	Bench	Cabinet	Column	mIoU
Static Corridor	0.92	0.92	0.92	0.82	–	–	0.92	0.90	0.83	0.89
Dynamic Corridor	0.94	0.86	0.91	0.83	–	–	0.91	0.93	0.86	0.89
Transition Space	0.91	–	0.79	–	0.92	0.85	–	–	–	0.87
Office	0.92	0.80	0.64	0.68	–	–	–	–	–	0.76

Table 2. Quantitative evaluation results (IoU and mIoU) across different indoor scenes in the custom dataset.

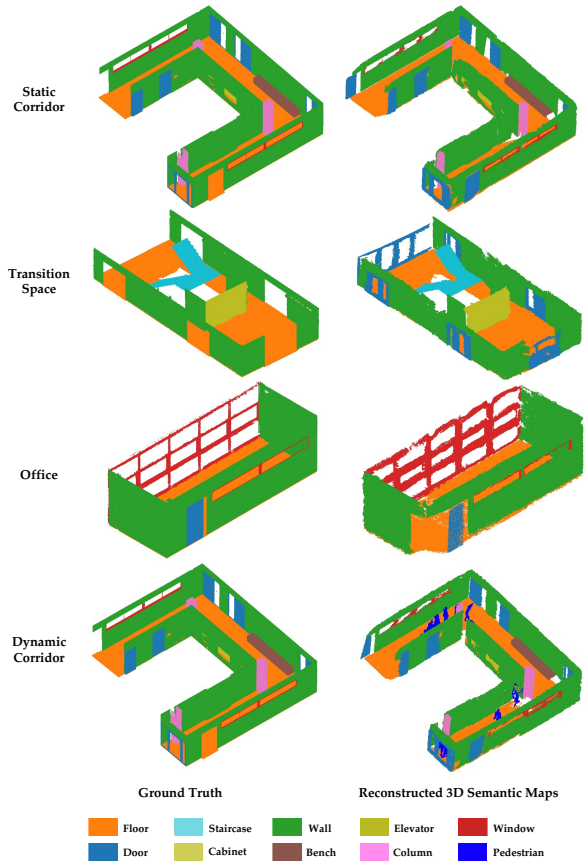

Figure 7. Qualitative comparison between reconstructed 3D semantic maps and ground-truth. Instances of the same category share a color while retaining unique instance IDs.

and 15.6 prompt points per prompt frame. This corresponds to an approximately fourfold reduction in overall prompt effort when using CPSS. Table 4 presents per-category IoU results under the same experimental settings as Table 3. Overall, CPSS achieves slightly higher segmentation accuracy, with improved structural completeness and fewer category-level misclassifications, particularly in regions such as cabinets, tables, and walls. These results, together with the qualitative examples in Figure 8, confirm that CPSS effectively balances segmentation accuracy and annotation efficiency across different environments.

5. Discussion

This work integrates RGB-D SLAM with SAM2 in a prompt-based framework to generate semantic 3D point clouds of indoor structural and functional elements. The proposed Category-Wise Prompt Segmentation Strategy (CPSS) reduces prompt interference and annotation effort while maintaining instance-level consistency.

Experiments show that the framework achieves centimeter-scale geometric consistency and high semantic consistency across various indoor environments using low-cost RGB-D sensors. Key structural components such as floors, walls, and doors are accurately reconstructed, and the transition space—with staircases and elevator areas—also maintains stable segmentation performance, indicating the robustness of CPSS in complex structural areas.

Evaluation on ScanNet further demonstrates the generalization capability of the approach. Across three representative scenes, CPSS achieves equal or higher mIoU than the multi-category prompting baseline while requiring substantially fewer correction frames and prompt points. Qualitative comparisons confirm cleaner instance boundaries and more coherent structural elements, occasionally exceeding ScanNet ground-truth boundary precision. On average, CPSS required only 1.2 prompt frames per instance and 4.7 prompt points per frame, reflecting a significant reduction in manual effort.

In the dynamic corridor sequence, segmentation of static structures remained accurate despite pedestrian motion. Keyframe selection avoids heavily occluded frames, and prompt-based control prevents moving objects from influencing the final map, demonstrating robustness to moderate indoor dynamics.

CPSS also reduces manual interaction across the custom dataset, typically requiring fewer than two keyframes and about five prompt points per instance. Combined with strong generalization on ScanNet, this highlights the scalability of the framework for large-scale or diverse indoor datasets.

Some scene-dependent limitations persist. The office environment shows lower mIoU primarily due to depth noise on reflective surfaces and reconstruction gaps from occlusions, suggesting that RGB-D mapping could benefit from improved depth filtering or sensor fusion.

Beyond segmentation accuracy, the resulting instance-level semantic point clouds support downstream applications such as efficient 3D annotation, Scan-to-BIM workflows, as illustrated in Figure 9, and integration into digital twin systems. By reducing annotation effort while preserving structural consistency, CPSS offers a practical and scalable solution for semantic 3D mapping in indoor environments.

6. Conclusion

This study presented a prompt-based 3D semantic mapping framework for instance-level segmentation of indoor structural and functional elements. The proposed Category-Wise Prompt Segmentation Strategy (CPSS) reduces cross-category interference and significantly lowers manual prompting effort while maintaining temporally consistent masks.

Experiments on custom RGB-D sequences and the ScanNet benchmark demonstrate centimeter-scale geometric consistency and strong semantic performance (mIoU up to 0.89 on

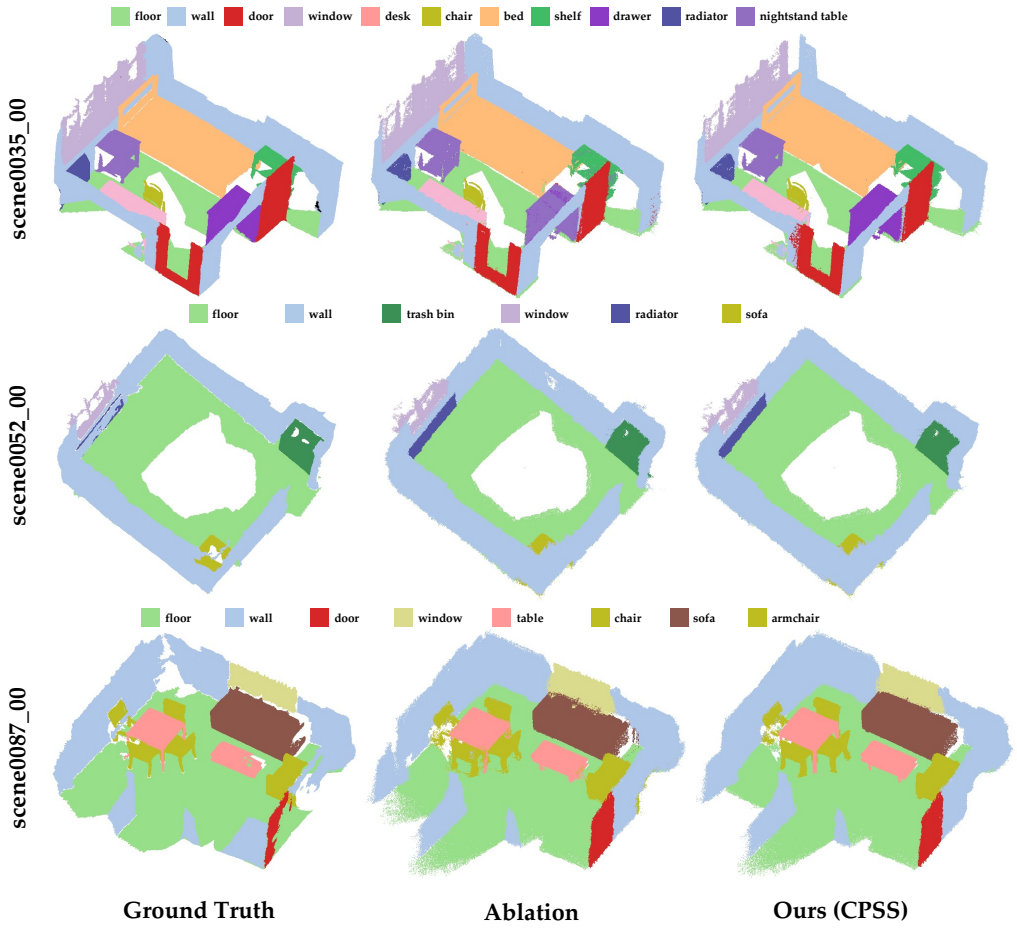

Figure 8. Qualitative comparison of 3D semantic mapping on selected ScanNet scenes, showing ground truth, ablation results, and the proposed CPSS. Instance colors follow ScanNet labels. CPSS produces cleaner segmentation with fewer prompts (see Tables 3 and 4).

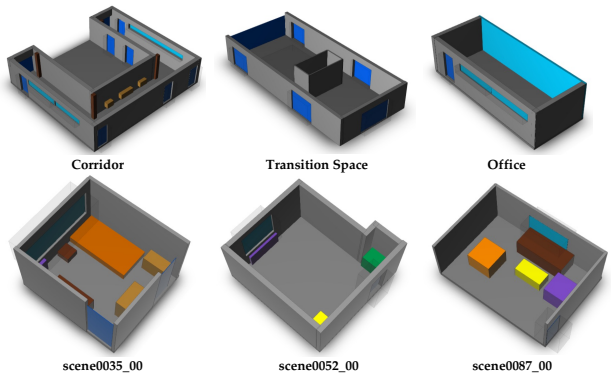

Figure 9. Lightweight BIM models reconstructed from the segmented semantic point clouds.

the custom dataset and 0.98 on ScanNet). Compared to multi-category prompting, CPSS achieves similar or improved per-class IoU with substantially fewer prompts.

The results confirm that accurate and semantically meaningful 3D maps can be constructed from RGB-D video data without retraining or environment-specific supervision, providing a practical solution for semantic indoor mapping tasks.

Future work will explore automatic prompt generation and further enhance robustness in more complex and large-scale environments, particularly under challenging sensing and scene conditions.

Acknowledgments

This research is supported by China Scholarship Council (CSC), Grant/Award Number: 202108130064.

References

Bescos, B., Facil, J. M., Civera, J., Neira, J., 2018. DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes. *IEEE Robot. Autom. Lett.*, 3(4), 4076–4083.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al., 2020. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 1877–1901.

Campos, C., Elvira, R., Rodriguez, J. J. G., M. Montiel, J. M., D. Tardos, J., 2021. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE Trans. Robot.*, 37(6), 1874–1890.

Chen, L., Ling, Z., Gao, Y. et al., 2023. A real-time semantic visual SLAM for dynamic environment based on deep learning

Scene ID	Setting	Category	Instances	Total/KeyFrames	Prompt Frames	Prompt Points	Correction Frames	Avg. Prompt Frames / Instance	Avg. Points / Prompt Frame
scene0035_00	Ablation (Baseline)	9	15	1475/286	17	196	12	1.13	11.53
scene0035_00	**CPSS (Improved)**	9	15	1475/286	**15**	**63**	**8**	**1.00**	**4.20**
scene0052_00	Ablation (Baseline)	6	9	991/229	16	210	9	1.78	13.13
scene0052_00	**CPSS (Improved)**	6	9	991/229	**14**	**88**	**8**	**1.56**	**6.29**
scene0087_00	Ablation (Baseline)	7	16	1396/359	18	399	7	1.13	22.17
scene0087_00	**CPSS (Improved)**	7	16	1396/359	**18**	**66**	**7**	**1.13**	**3.67**

Table 3. Quantitative comparison of the baseline ablation and the proposed CPSS, showing reduced prompting effort with similar correction requirements and keyframe coverage. Prompt frames denote keyframes in which manual point prompts are provided.

Scene ID	Setting	Floor	Wall	Window	Trash Bin	Radiator	Sofa	Bed	Door	Cabinet	Table	Chair	mIoU
scene0035_00	Ablation (Baseline)	0.99	0.95	0.98	–	0.81	–	0.99	0.86	0.84	0.56	0.96	0.88
scene0035_00	**CPSS (Improved)**	0.99	0.93	0.99	–	0.99	–	0.99	0.96	0.99	0.98	0.97	**0.98**
scene0052_00	Ablation (Baseline)	0.98	0.88	0.82	0.99	0.49	0.88	–	–	–	–	–	0.84
scene0052_00	**CPSS (Improved)**	0.98	0.90	0.82	0.99	0.49	0.88	–	–	–	–	–	**0.85**
scene0087_00	Ablation (Baseline)	0.98	0.53	0.85	–	–	0.97	–	0.38	–	0.92	0.97	0.80
scene0087_00	**CPSS (Improved)**	0.97	0.57	0.87	–	–	0.99	–	0.36	–	0.96	0.98	**0.81**

Table 4. Quantitative comparison of per-category IoU between the proposed CPSS and multi-category prompting (Ablation) on selected ScanNet scenes.

and dynamic probabilistic propagation. *Complex Intell. Syst.*, 9, 5653–5677.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017a. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2432–2443.

Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C., 2017b. BundleFusion: Real-Time Globally Consistent 3D Reconstruction Using On-the-Fly Surface Reintegration. *ACM Trans. Graph.*, 36(3).

Guo, Z., Zhang, R., Zhu, X., Tong, C., Gao, P., Li, C., Heng, P.-A., 2024. Sam2point: Segment any 3d as videos in zero-shot and promptable manners. preprint.

He, Y., Chen, B., Motagh, M., Zhu, Y., Shao, S., Li, J., Zhang, B., Kaufmann, H., 2025. Zero-shot detection for InSAR-based land displacement by the deformation-prompt-based SAM method. *Int. J. Appl. Earth Obs.*, 136, 104407.

Hou, J., Goebel, M., Hübner, P., Iwaszczuk, D., 2023. OCTREE-BASED APPROACH FOR REAL-TIME 3D INDOOR MAPPING USING RGB-D VIDEO DATA. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-1/W1-2023, 183–190.

Hübner, P., Hou, J., Iwaszczuk, D., 2023. Evaluation of Intel RealSense D455 Camera Depth Estimation for Indoor SLAM Applications. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-1/W2-2023, 1207–1214.

Kang, J., Chen, N., Li, M., Mao, S., Zhang, H., Fan, Y., Liu, H., 2024. A Point Cloud Segmentation Method for Dim and Cluttered Underground Tunnel Scenes Based on the Segment Anything Model. *Remote Sens.*, 16(1).

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment anything. *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 3992–4003.

Landrieu, L., Simonovsky, M., 2018. Large-scale point cloud semantic segmentation with superpoint graphs. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 4558–4567.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 652–660.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017b. Pointnet++: deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, NIPS’17, Curran Associates Inc., Red Hook, NY, USA, 5105–5114.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C., 2024. Sam 2: Segment anything in images and videos. preprint.

Wang, J., Liu, Z., Zhao, L., Wu, Z., Ma, C., Yu, S., Dai, H., Yang, Q., Liu, Y., Zhang, S., Shi, E., Pan, Y., Zhang, T., Zhu, D., Li, X., Jiang, X., Ge, B., Yuan, Y., Shen, D., Liu, T., Zhang, S., 2023. Review of large vision models and visual prompt engineering. *Meta Radiol*, 1(3), 100047.

Wang, Z., Yu, X., Rao, Y., Zhou, J., Lu, J., 2022. P2p: tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, NIPS ’22, Curran Associates Inc., Red Hook, NY, USA, 14388 – 14402.

Xie, S., Gu, J., Guo, D., Qi, C. R., Guibas, L. J., Litany, O., 2020. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Lecture Notes in Computer Science, 12348, Springer, 574–591.

Xu, X., Lee, G. H., 2020. Weakly supervised semantic point cloud segmentation: Towards 10× fewer labels. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 13703–13712.

Yang, Y., Wu, X., He, T., Zhao, H., Liu, X., 2023. Sam3d: Segment anything in 3d scenes.

Yarovoi, A., Cho, Y. K., 2024. Review of simultaneous localization and mapping (SLAM) for construction robotics applications. *Automat. Constr.*, 162, 105344.

Yin, Y., Liu, Y., Xiao, Y., Cohen-Or, D., Huang, J., Chen, B., 2024. Sai3d: Segment any instance in 3d scenes. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 3292–3302.