Unified Promptable Panoptic Mapping with Dynamic Labeling Using Foundation Models

Mohamad Al Mdfaa *, Raghad Salameh, Geesara Kulathunga *, Sergey Zagoruyko, Gonzalo Ferrer

Posted Date: 1 December 2025

doi: 10.20944/preprints202512.0137.v1

Keywords: semantic scene understanding; panoptic mapping; open-vocabulary panoptic segmentation; foundation models; vision language models; robotics

 doi:10.20944/preprints202512.0137.v1

Article

Unified Promptable Panoptic Mapping with Dynamic Labeling Using Foundation Models

Mohamad Al Mdfaa [1,*], Raghad Salameh [1], Geesara Kulathunga [2], Sergey Zagoruyko [3] and Gonzalo Ferrer [1]

[1] Applied AI Institute, Moscow, Russia
[2] School of Computer Science, University of Lincoln, United Kingdom
[3] Independent Researcher
* Correspondence: m.aimdfaa@applied-ai.ru

Abstract

Panoptic maps enable robots to reason about both geometry and semantics. However, open-vocabulary models repeatedly produce closely related labels that split panoptic entities and degrade volumetric consistency. The proposed UPPM advances open-world scene understanding by leveraging foundation models to introduce a panoptic Dynamic Descriptor that reconciles open-vocabulary labels with unified category structure and geometric size priors. The fusion for such dynamic descriptors is performed within a multi-resolution multi-TSDF map using language-guided open-vocabulary panoptic segmentation, and semantic retrieval, resulting in a persistent and promptable panoptic map without additional model training. Based on our evaluation experiments, UPPM shows the best overall performance in terms of the map reconstruction accuracy and the panoptic segmentation quality. The ablation study investigates the contribution for each component of UPPM (custom NMS, blurry-frame filtering, and unified semantics) to the overall system performance. Consequently, UPPM preserves open-vocabulary interpretability while delivering strong geometric and panoptic accuracy. The code will be released upon acceptance.

Keywords: semantic scene understanding; panoptic mapping; open-vocabulary panoptic segmentation; foundation models; vision language models; robotics

1. Introduction

Panoptic mapping connects 3D geometry with semantics so that robots can localize, plan, and query in structured environments. Systems such as SemanticFusion [1], SLAM++ [2], Panoptic Fusion [3] and DKB-SLAM [4] deliver spatial consistency but assume a predefined set of categories; open-vocabulary labels therefore fragment instances or fall back to generic labels. Recent advances in high-capacity vision segmentation [5–8] and open-vocabulary or vision-language segmentation [9–13] broaden the 2D label space, but their predictions still live on individual images; they do not resolve how to reconcile those labels with volumetric priors, maintain cross-view consistency, or reason about object scale in 3D.

A wave of recent works extend open-vocabulary reasoning into 3D. Panoptic Vision-Language Feature Fields (PVLFF) learn a radiance field together with semantic and instance feature heads by distilling vision-language embeddings into a feature field and clustering the learned instance features [14]. Despite achieving open-vocabulary panoptic segmentation, PVLFF must be optimized offline for each scene [14]. OpenVox introduces a real-time, incremental open-vocabulary voxel representation that uses caption-enhanced detection and probabilistic instance voxels; its incremental fusion decomposes into instance association and live map evolution to increase robustness [15]. While effective, OpenVox focuses on instance association rather than unifying synonyms or transferring size priors.

 doi:10.20944/preprints202512.0137.v1

Other open-vocabulary mapping frameworks embed language features in 3D without panoptic constraints. ConceptFusion [16] projects pixel-aligned CLIP [17] and DINO [18] features into 3D to enable multimodal queries; the resulting pixel-aligned embeddings capture fine-grained, long-tailed concepts but do not preserve instance continuity or temporal consistency. ConceptGraphs [19] associates class-agnostic masks across views to build an object-centric 3D scene graph and employs a large language model to caption each object and infer spatial relationships; this supports complex language queries but lacks volumetric geometry and per-voxel size priors. Clio [20] formulates task-driven scene understanding through the information bottleneck, clustering 3D primitives into task-relevant objects and building a hierarchical scene graph, yet its segmentation granularity depends on task-specific thresholds. OpenScene [21] trains a 3D network to co-embed each point with text and image features in the CLIP space [17], enabling zero-shot queries of materials, affordances and room types, but its dense per-point features do not distinguish individual instances.

Unified Promptable Panoptic Mapping (UPPM) addresses these limitations. We use off-the-shelf vision-language models [22] to generate elementary descriptors, where each object elementary descriptor conveys a rich open-vocabulary label and then perform semantic retrieval to attach it to a unified category with an inherited volumetric size prior. Rather than treating these elementary descriptors in isolation, UPPM groups them within the unified panoptic fusion block into a dynamic descriptor for each object. Each dynamic descriptor stores the aggregated original captions, the assigned unified category and a size prior for a single object. Each segmentation mask with its elementary descriptor are associated to the submap (i.e. 3D panoptic entity which could be an object, a piece of background or free space) that has the highest Intersection over Union (IoU) between predicted and rendered mask and the same class label. To avoid spurious associations, a minimum IoU of $\xi_{IoU} = 0.1$ is required (Figures 1 and 3). UPPM therefore preserves the expressiveness of open-vocabulary prompts while producing category-consistent panoptic maps without per-scene training or graph-based reasoning. The unified descriptors also enable promptable downstream queries, such as language-conditioned object retrieval, without compromising geometric accuracy. Fig 2 showcases how the resulting map enables language-conditioned retrieval without additional training.

We assess UPPM on three evaluation regimes: Segmentation-to-Map evaluates the impact of 2D predictions on the quality of 3D reconstruction. Map-to-Map directly compares reconstructed 3D maps with ground-truth geometry, while Segmentation-to-Segmentation evaluates 2D panoptic segmentation performance.

This work has the following contributions:

(i) We introduce a panoptic dynamic descriptor for each unique object that ties together the open-vocabulary elementary descriptors associated with that object across frames. Rather than assigning labels per detection, UPPM aggregates the object open-vocabulary labels produced by vision-language models [22] and post-processed by part-of-speech tagging [23], maps them to a single unified semantic category and size prior, and stores them as the dynamic descriptor (Sections 3.2 and 3.2.2).

(ii) We build the UPPM pipeline around these dynamic descriptors to produce 3D panoptic maps that are geometrically accurate, semantically consistent, and naturally queryable in language (Section 3).

(iii) We extensively evaluate UPPM across Segmentation-to-Map, Map-to-Map, and Segmentation-to-Segmentation regimes on three different datasets (ScanNet v2, RIO, and Flat), and perform ablations of unified semantics, custom NMS, blurry-frame filtering, and tag usage (Sections 4.1–4.3, 4.5, 4.5.1 and 4.5.2). Together, these experiments show that dynamic descriptors improve both geometric reconstruction and panoptic quality while enabling downstream language-conditioned tasks such as object retrieval.

 doi:10.20944/preprints202512.0137.v1

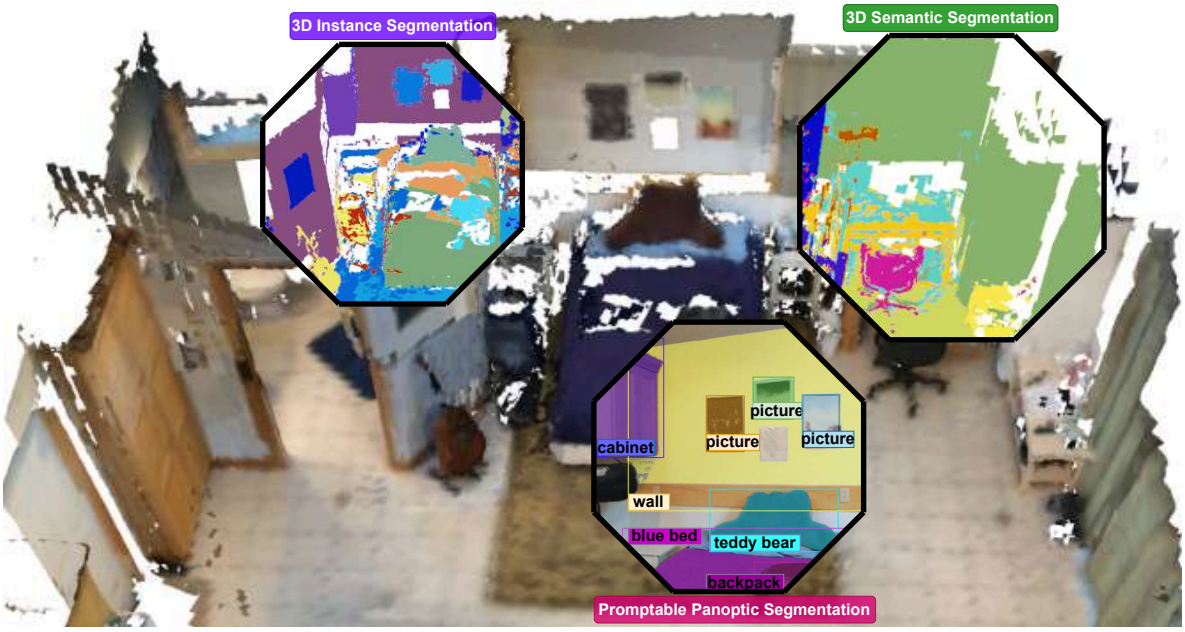

Figure 1. Overview of Unified Promptable Panoptic Mapping (UPPM). Dynamic descriptors aggregate open-vocabulary labels, unified semantic categories, and size priors, and are fused into a multi-resolution multi-TSDF map that preserves panoptic consistency while remaining queryable with natural-language prompts.

2. Related Work

Recent semantic mapping systems span multi-resolution volumetric fusion and metric-semantic SLAM. Panoptic Multi-TSDFs [24] and the hierarchical DHP-Mapping framework [25] maintain panoptic submaps with label optimization, while Fusion++ [26] instantiates object-level volumetric models and Kimera [27] produces metric-semantic reconstructions and dynamic scene graphs. Likewise, methods such as [28] explore semantic segmentation and room-layout cues to improve robustness. Despite strong performance, these pipelines typically operate with closed vocabularies, which is limited in the real-world applications.

2.1. Open-Vocabulary 3D Mapping

To relax vocabulary constraints, vision-language extensions inject language supervision into volumetric pipelines. PVLFF [14] learns a radiance field with semantic and instance feature fields and performs open-vocabulary panoptic segmentation by clustering instance features; however, it requires per-scene offline training and is not designed for real-time incremental mapping. OpenVox [15] introduces an incremental instance-level open-vocabulary mapping framework with caption-enhanced detection and a probabilistic voxel representation. It maintains an embedding codebook for each instance and emphasizes robust association and mapping, rather than unifying synonyms into a common category or transferring volumetric size priors. In contrast, UPPM leverages off-the-shelf foundation models without additional training, adds a semantic retrieval stage to assign each object a unified category and size, and fuses dynamic descriptors into a multi-resolution multi-TSDF map.

A number of recent works further expand open-vocabulary mapping by exploring new architectures and representations. FindAnything [29] integrates vision-language features into volumetric occupancy submaps to support object-centric mapping with natural-language queries on lightweight platforms. OVO [30] employs online detection and tracking of 3D segments with CLIP-based descriptor fusion to achieve open-vocabulary mapping within a SLAM. RAZER [31] combines GPU-accelerated TSDF reconstruction with hierarchical association and spatio-temporal aggregation to produce zero-shot panoptic maps, reporting real-time performance on modern GPUs. OpenGS-Fusion [32] couples 3D Gaussian splats with TSDFs and introduces language-guided thresholding to improve volumetric segmentation quality. DualMap [33] uses a dual representation (abstract/global and concrete/local) and a hybrid segmentation front end to handle dynamic scenes and language-conditioned navigation.

 doi:10.20944/preprints202512.0137.v1

MR-COGraphs [34] compresses semantic features in scene graphs for communication-efficient multi-robot mapping. 3D-AVS [35] proposes automatic vocabulary discovery for LiDAR point clouds, and Open-Vocabulary Functional 3D Scene Graphs [36] encode objects, interactive elements, and functional relations using vision-language and large language models. These methods illustrate the community's pivot toward open-vocabulary and hierarchical understanding, but many rely on neural radiance fields, Gaussian splats, or graph reasoning and do not unify synonyms or transfer volumetric size priors during mapping.

2.2. Open-Set Scene Graphs and Feature Maps

OpenScene [21] computes CLIP-aligned features for every 3D point using multi-view 2D fusion and 3D distillation to enable task-agnostic open-vocabulary retrieval; it constructs a semantic map but does not perform instance-level segmentation. Beyond OpenScene, scene-graph and radiance-field approaches such as Clio [20], Concept-Graphs [19], LangSplat [37], Bayesian-Fields [38], and OpenGaussian [39] enable language-guided reasoning, yet often trade off panoptic completeness or require per-query optimization. HOV-SG [40] builds hierarchical open-vocabulary 3D scene graphs by extracting CLIP embeddings per SAM mask and back-projecting fused features to 3D; unlike UPPM, it does not normalize embeddings into a fixed unified category space with inherited size priors, so it does not directly yield category-consistent panoptic instance maps. Consistent with semantics-conditioned scale choices in panoptic mapping [24,26], UPPM maintains a coarse size prior (Small/Medium/Large) during mapping.

2.3. Foundation Vision-Language Models

Foundation models [41] provide the building blocks for open-world perception. Multimodal encoders such as CLIP [17] supply open-vocabulary embeddings; promptable segmentation models like Segment Anything [42] yield category-agnostic masks; and multimodal assistants including LLaVA [43] generate rich scene descriptions that can seed dynamic descriptors.

ConceptFusion [16], OVIR-3D [44], and OpenMask3D [45] handle open-vocabulary cues in 3D but lack a shared categorical structure. UPPM combines dynamic descriptors with unified semantics so that the map remains panoptic while supporting promptable queries and coarse-to-fine relational reasoning within a single volumetric representation.

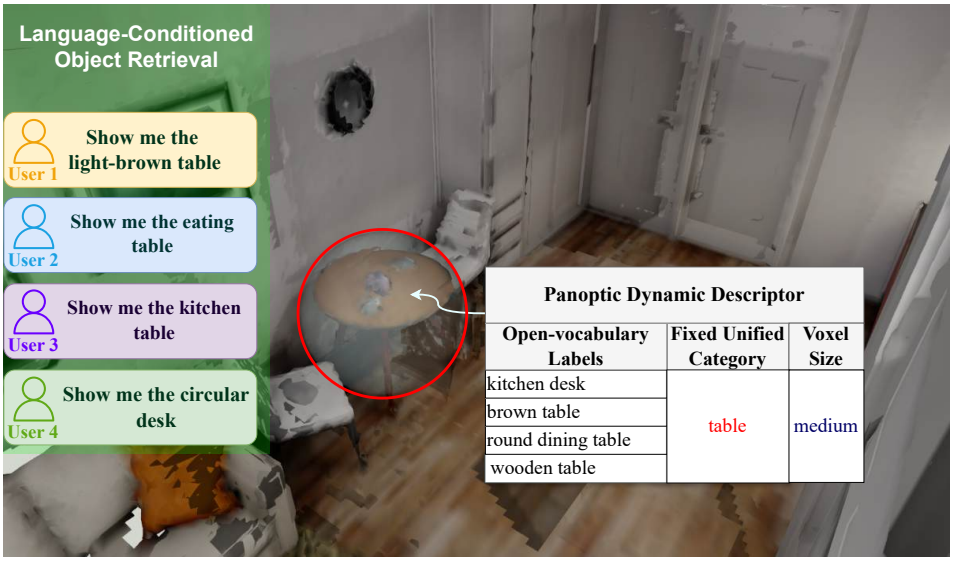

Figure 2. Language-conditioned object retrieval via dynamic descriptors. Different textual queries consistently retrieve the same 3D object, as each submap stores aggregated open-vocabulary labels, a unified category, and a size prior.

 doi:10.20944/preprints202512.0137.v1

3. Method

UPPM addresses the visual semantic mapping problem by utilizing foundation models [22,42,46–48] to construct dynamic descriptors that help to build precise reconstructions with rich semantics. The term 'dynamic descriptor' here refers to the object-level data structure formed by aggregating all open-vocabulary cues observed for the same object over time, a shared unified category and size prior, an aggregation that is performed during the unified panoptic fusion stage. The mapping pipeline follows a systematic multi-stage process: First, with the help of vision-language models, we extract rich open-vocabulary labels from the RGB inputs. Second, a semantic retrieval module attaches to these labels a unified category with the associated volumetric size priors, yielding elementary descriptors for each label. Third, language-conditioned open-vocabulary object detection and promptable image segmentation are guided by these elementary descriptors. Finally, during the unified panoptic fusion stage the elementary descriptors associated with each object are aggregated into a single dynamic descriptor. Consequently, the resulting semantic and instance segmentations with the dynamic descriptors are fused into a unified promptable panoptic map (UPPM) that is both geometrically accurate and semantically consistent. The complete mapping pipeline is shown in Figure 3.

3.1. Problem Formulation

Given a sequence of posed RGBD images $\mathcal{U} = \{U_1, \ldots, U_T\}$ where each $U_t = (I_t, D_t, P_t)$ consists of an RGB image $I_t \in \mathbb{R}^{H \times W \times 3}$, a depth image $D_t \in \mathbb{R}^{H \times W}$, and a camera pose $P_t \in SE(3)$, our goal is to construct a unified promptable panoptic map $\mathcal{M}$ that satisfies the following objectives:

1. **Geometric Reconstruction:** Generate an accurate 3D reconstruction of the environment represented as a multi-resolution multi-TSDF map $\mathcal{M}_g$.
2. **Dynamic Labeling with Unified Semantics:** Construct the set of dynamic descriptors $\mathcal{M}_d$ by first generating elementary descriptors from open-vocabulary cues and then, during the unified panoptic fusion stage, aggregating all elementary descriptors that refer to the same object across frames into a single dynamic descriptor.

Specifically, assign a dynamic descriptor $d_i \in \mathcal{M}_d$ to each unique object $o_i \in \mathcal{O}$, where $\mathcal{O}$ is the set of all objects in the scene. Each dynamic descriptor collects all open-vocabulary labels that have been associated with that object across frames and consolidates them under a single unified category and size prior. Formally, $d_i = \langle O_i, \hat{c}_i, \hat{s}_i \rangle$ where O_i is the set of open-vocabulary labels aggregated for object o_i, $\hat{c}_i \in \mathcal{C}$ is the unified category chosen through (1), and $\hat{s}_i$ is the inherited size prior. We choose the COCO-Stuff classes [49] as the fixed unified categories due to their comprehensive coverage of common indoor objects and well-established hierarchy, providing a robust foundation for semantic mapping tasks. The final output is a unified promptable panoptic map $\mathcal{M} = (\mathcal{M}_g \bigoplus \mathcal{M}_d)$, where the fusion operator $\bigoplus$ combines geometric reconstruction $\mathcal{M}_g$ with dynamic descriptors $\mathcal{M}_d$ into a unified volumetric representation integrating spatial and semantic information.

3.2. Dynamic Labeling and Segmentation

As shown in Figure 3, the per-frame processing pipeline consists of three modules that construct structured dynamic descriptors from RGB images. Visual-Linguistic Feature Extraction++ (VLFE++) generates open-vocabulary labels, Semantic Retrieval (SR) maps such labels to unified categories, and the open-vocabulary panoptic segmentation module consumes these labels to produce semantic and instance segmentations that are ready for fusion.

3.2.1. Visual-Linguistic Feature Extraction++ (VLFE++)

VLFE++ leverages vision-language models to gather open-vocabulary labels from visual input. As shown in Figure 3, VLFE++ applies a three-step extraction procedure:

- **Visual Linguistic Features Extraction (VLFE):** a pre-trained vision-language model [22] is employed to produce captions and tags that cover rich semantic descriptions for subsequent processing steps.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2025 doi:10.20944/preprints202512.0137.v1

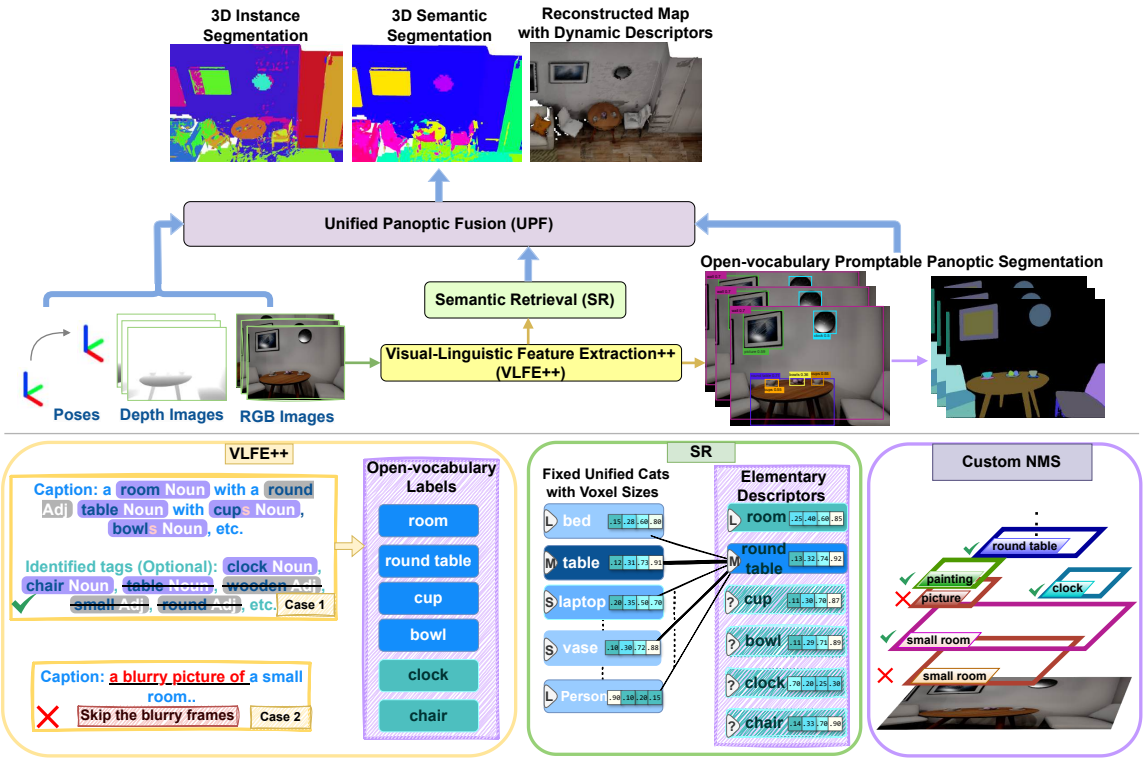

Figure 3. System overview of Unified Promptable Panoptic Mapping (UPPM). RGB frames are processed by VLFE++ to generate caption- and tag-based open-vocabulary labels. Semantic Retrieval (SR) maps these labels to unified semantic categories and size priors, forming elementary descriptors $e_i = \langle o_i, E(o_i), \hat{c}_i, \hat{s}_i \rangle$. Open-vocabulary promptable panoptic segmentation consume these descriptors, together with custom NMS, to produce semantic and instance segmentations. Unified panoptic fusion then fuses geometry and elementary descriptors into a multi-resolution multi-TSDF representation with one dynamic descriptor per object, producing the unified promptable panoptic map (UPPM) $\mathcal{M} = (\mathcal{M}_g \oplus \mathcal{M}_d)$.

- **Part-Of-Speech (POS) Tagging:** A single-layer perceptron network [23] is applied to extract nouns, noun phrases, and modifiers. The resulting candidate list is passed to the next step.
- **Lemmatization:** Lemmatization [50] is used to map the derived forms from the candidate list to their normalized base forms (e.g., 'apples' → 'apple'), providing canonical labels for retrieval.

The generated open-vocabulary labels serve as input for the next pipeline component, SR.

3.2.2. Semantic Retrieval (SR)

SR maps open-vocabulary labels to unified categories while preserving these labels. Let $\mathcal{O}_{vocab}$ be the set of open-vocabulary labels from VLFE++ and $\mathcal{C}$ be the set of fixed unified categories, where each class $(c_i \in \mathcal{C})$ has a corresponding voxel size attribute (v_i), which takes one of these empirical scale priors - small $(v = 2cm)$, medium $(v = 3cm)$, or large $(v = 5cm)$ that were introduced in PanMap [24]. This choice maintains compatibility with established panoptic mapping pipelines and ensure consistent comparison. We also keep $v_{freespace} = 30cm$ for all settings. Our goal is to determine the most appropriate unified category $\hat{c}$ for any given open-vocabulary label $o \in \mathcal{O}_{vocab}$:

$$\hat{c} = \arg\max_{c \in \mathcal{C}} \cos(E(o), E(c)), \tag{1}$$

where $\cos(\cdot)$ is the cosine similarity between embeddings and $E(\cdot)$ denotes the embedding function. The predicted size $\hat{v} = v_{\hat{c}}$ is inherited from the assigned unified category $\hat{c}$. Following Panoptic Multi-TSDFs [24], each unified category carries a manually curated volumetric prior; UPPM transfers this size to new elementary descriptors through SR so that dynamic descriptors remain metrically

 doi:10.20944/preprints202512.0137.v1

grounded. We employ MPNet [47] for embedding generation and perform semantic search [51] using cosine similarity to select the closest unified category. For each open-vocabulary label $o_i \in \mathcal{O}_{\text{vocab}}$, SR finds the corresponding unified category $\hat{c}$ and predicts its size $\hat{v}$. These assignments yield an elementary descriptor $e_i = \langle o_i, \hat{c}_i, \hat{v}_i \rangle$ for the label. During the unified panoptic fusion stage these elementary descriptors are aggregated for each object into a single dynamic descriptor that collects all of its labels and shares a unified category and size prior. The elementary descriptor for a label is defined as:

$$e_i = \langle o_i, \hat{c}_i, \hat{v}_i \rangle, \tag{2}$$

where o_i is the open-vocabulary label from VLFE++, $\hat{c}_i$ is the unified category assigned by (1), and $\hat{v}_i$ is the inherited size prior. These elementary descriptors drive the open-vocabulary promptable panoptic segmentation and later serve as building blocks for the object-level dynamic descriptors formed in the unified panoptic fusion stage (Figure 2).

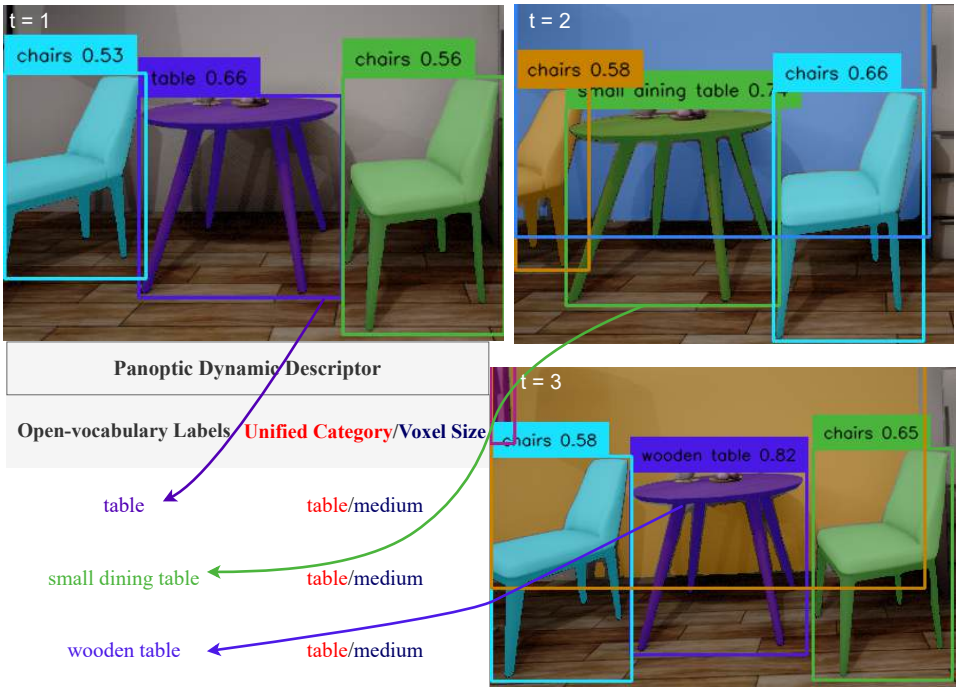

Figure 4. Panoptic dynamic descriptors unify open-vocabulary labels over the time. As observations accumulate across frames, the descriptor remains attached to the same object in the panoptic map, preserving promptable yet category-consistent semantics.

3.2.3. Open-vocabulary Panoptic Segmentation

UPPM employs Grounding-DINO [46] as the object detector, leveraging its capability to process curated open-vocabulary labels. Grounding-DINO generates precise bounding boxes that serve as input prompts for the Segment Anything Model (SAM) [42], enabling the creation of high-quality segmentation for each detected object in the scene. Furthermore, as shown in Table 5, a custom Non-Maximum Suppression (NMS) technique is proposed to address the open-vocabulary detection challenges [46]. While traditional NMS approaches like per-class NMS [52] remove redundant bounding boxes based on Intersection-over-Union (IoU) and confidence scores within the same class, our custom NMS handles two additional scenarios:

1. **Prioritizing caption-derived labels:** When overlapping bounding boxes have different labels but belong to the same unified category, we prioritize the box whose label originates from the image caption over those derived from the identified tags, as shown in Figure 3. This context-aware

 doi:10.20944/preprints202512.0137.v1

selection leverages the richer semantic information provided by captions compared to identified tags [22].

2. **Standard duplicate removal:** For overlapping bounding boxes b_a, b_b with identical labels where $\text{IoU}(b_a, b_b) \geq \tau$, we apply conventional suppression based on confidence scores, retaining the box with higher detection confidence.

3.3. Unified Panoptic Fusion (UPF)

Building upon the open-vocabulary panoptic segmentation results, we integrate these outputs into a comprehensive 3D representation by adopting a panoptic multi-TSDFs-based approach [24]. Traditional panoptic mapping approaches rely on fixed semantic categories, limiting their ability to handle novel objects. Our Unified Panoptic Fusion extends beyond these limitations by integrating dynamic descriptors while maintaining semantic consistency through unified categorization. We build upon the object-centric mapping framework introduced in [24], specifically leveraging its submap-based architecture. Submaps are localized volumetric representations that divide large-scale environments into smaller, manageable parts, where each submap contains geometric and semantic data including panoptic information, object instances, and semantic categories, along with transformation and tracking data. These submaps enable efficient processing of large-scale environments while significantly reducing computational complexity. UPF treats each object, background region, or free space as its own submap stored in a separate TSDF grid, with its own instance identity and semantic class. The rendered masks are associated with panoptic segmentations based on Intersection-over-Union (IoU). The new observation is fused into the submap only if the IoU exceeds a small threshold; otherwise, a new submap is created. If μ_e denotes the mask for the elementary descriptor e in the new frame, and ρ_s denotes the rendered mask of a submap s, then the assigned submap s^* is as shown in Equation (3).

$$s^*(e) = \begin{cases} \arg\max_{s \in \mathcal{S}} \text{IoU}(\mu_e, \rho_s), & \text{if } \max_{s \in \mathcal{S}} \text{IoU}(\mu_e, \rho_s) \geq \xi_{\text{IoU}}, \\ s_{new}, & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathcal{S}$ is the set of submaps, s_{new} is a new submap and ξ_{IoU} is a small association threshold.

Our key contribution is to enrich these submaps with dynamic descriptors, as illustrated in Figure 4. Whereas PanMap [24] assign a fixed semantic label to every submap, UPPM augments each submap with a dynamic descriptor that aggregates all open-vocabulary labels observed for that object across frames, along with the retrieved unified category and inherited voxel size prior. As new frames are fused, their elementary descriptors are merged into the submap so that closely related elementary descriptors accumulate into a single consolidated dynamic descriptor. These descriptors travel with the submap through the association and change-detection machinery described above, ensuring that there is exactly one descriptor per object rather than per detection. This grouping enables flexible open-vocabulary querying without sacrificing the hard panoptic consistency afforded by the underlying PanMap framework. The Unified Panoptic Fusion component takes as input posed RGBD data, dynamic descriptors, detected objects, and panoptic segmentation, reconstructing the Unified Promptable Panoptic Map (UPPM) $\mathcal{M}$ by integrating this information into the enhanced submap structure. The resulting map is accurately reconstructed and enriched with dynamic descriptors, facilitating downstream tasks such as navigation and localization.

4. Results

We evaluate UPPM across three regimes: **Segmentation-to-Map** measures how 2D panoptic segmentations drive 3D reconstruction when PanMap [24] is paired with different segmentation backbones across ScanNet v2 [53], RIO [54], and Flat [24]. Metrics include geometric accuracy, completeness, Hausdorff distance, completion ratio, Chamfer distance, fine-detail F1-score, and runtime. **Map-to-Map** compares reconstructed maps directly against the ground-truth geometry using accuracy, completeness, and Chamfer-L1. **Segmentation-to-Segmentation** evaluates panoptic segmentation

 doi:10.20944/preprints202512.0137.v1

quality on Flat using PQ, RQ, SQ, and mAP metrics. Experiments ran on a server with an NVIDIA A100 (80 GB) GPU and 256 GB RAM.

4.1. Segmentation-to-Map Evaluation

The impact of segmentation quality on geometric reconstruction is assessed by comparing UPPM with PanMap [24], each deployed with different segmentation backbones: MaskDINO [5], OpenSeeD [9], OMG-Seg [10], and ground-truth segmentation (GTS).

UPPM achieves strong performance across multiple metrics (Table 1). The proposed method attains optimal accuracy on RIO (1.55 *cm*) and Flat (0.61 *cm*) with competitive performance on ScanNet v2 (3.30 *cm*), where *Accuracy* measures how close reconstructed points are to ground truth, calculated as mean absolute error from reconstruction (R) to ground truth (G). This improved accuracy on RIO shows UPPM's robustness to motion blur and sensor noise common in real-world datasets. To assess how thoroughly the reconstruction captures the ground truth scene, *Completeness* is defined as the mean absolute error from ground truth (G) to reconstruction (R). UPPM attains outstanding completeness on Flat (1.10 *cm*) and competitive performance on ScanNet v2 (1.78 *cm*) and RIO (1.19 *cm*). This pattern reflects the influence of dataset characteristics: Flat [24], being simulated, enables precise reconstruction and higher completeness, while the real-world conditions and noise in RIO [54] and ScanNet v2 [53] make achieving high completeness more challenging. The metrics Accuracy and Completeness are asymmetrical, as distances 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 Accuracy component suggests potential inaccuracies in reconstruction, while a larger Completeness component suggests potential incompleteness.

To estimate how completely the reconstruction covers the ground truth, we use the *Completion Ratio*, defined as the percentage of observed ground truth points reconstructed within a 5 *cm* threshold. UPPM achieves strong completion ratios (81.42% on ScanNet v2, 74.14% on RIO, 70.76% on Flat). The higher completion ratio on ScanNet v2 reflects the ability to leverage higher quality data and more diverse semantic categories, while the lower ratio on RIO indicates the challenge of noisy real-world data.

To quantify geometric similarity between the reconstructed and ground truth point clouds, *Chamfer distance* $d_{\text{ch}}(G, R)$ is used. This metric provides an overall assessment of how closely the reconstructed and ground truth point clouds match:

$$d_{\text{ch}}(G,R) = \underbrace{\sum_{g \in G} \min_{r \in R} \|g - r\|_2^2}_{\text{Completeness}} + \underbrace{\sum_{r \in R} \min_{g \in G} \|g - r\|_2^2}_{\text{Accuracy}} \tag{4}$$

UPPM achieves exceptional Chamfer distance on RIO and competitive performance on ScanNet v2 and Flat (ScanNet v2: 12.64 *cm*, RIO: 5.65 *cm*, Flat: 2.64 *cm*). This lower Chamfer distance reflects the effectiveness of the multi-resolution Multi-TSDF approach, which balances accuracy and completeness. The substantial improvements on RIO indicate UPPM's robustness to motion blur and sensor noise, while competitive performance on ScanNet v2 shows its ability to handle complex real-world scenes with diverse semantic categories.

To estimate the worst-case reconstruction error, we use the *Hausdorff distance* (d_H), as defined in Equation (5). This metric measures the greatest distance from a point in one set to the closest point in the other set:

$$d_H(G,R) = \max\left(\max_{g \in G} \min_{r \in R} \|g - r\|_2, \max_{r \in R} \min_{g \in G} \|r - g\|_2 \right) \tag{5}$$

UPPM achieves competitive Hausdorff distance performance with optimal performance on RIO (10.05 cm vs 18.55-21.60 cm from competitors), demonstrating UPPM's robustness in minimizing worst-case reconstruction errors. The Hausdorff distance on Flat (4.75 *cm*) highlights how the controlled environment helps minimize outlier errors. The strong performance on ScanNet v2 (25.53 *cm*) demon-

 doi:10.20944/preprints202512.0137.v1

strates UPPM's ability to manage complex scenes with varied geometric structures while keeping worst-case errors within reasonable limits.

To measure the ability to reconstruct fine-grained details while accounting for both false positives and false negatives, *F1-Score* is calculated after removing background classes (e.g., floor, ceiling, wall). UPPM achieves 2.27 FPS on Flat (batch size 1), demonstrating computational efficiency suitable for real-time applications. UPPM attains the highest F1-score (79.54%), indicating better fine-detail reconstruction. This result is enabled by the custom NMS strategy, which removes duplicates while preserving details, and by dynamic labeling that improves coverage without over-segmentation.

Figure 5. Qualitative comparison of map reconstruction on the Flat dataset [24]. A MaskDINO-based baseline attains high completion but misses fine-grained structures such as wall decorations, whereas UPPM recovers these details while preserving large-scale geometry.

The qualitative analysis reveals important insights about reconstruction quality. As shown in Figure 5, while MaskDINO achieves better completion ratio on some datasets, it demonstrates better reconstruction for large structures like walls and floors but fails to capture intricate details such as paintings on walls, which hold significant semantic value. Conversely, UPPM successfully reconstructs these fine-grained elements, demonstrating its high capability in detailed scene understanding.

Figure 6. Qualitative comparison on the RIO dataset [54]. **Top:** Ground-truth panoptic segmentation and corresponding 3D map. **Bottom:** UPPM output with sharper object boundaries and more coherent semantic regions in cluttered, real-world scenes.

 doi:10.20944/preprints202512.0137.v1

Table 1. Quantitative evaluation across all datasets. All baselines use the PanMap framework [24] with different segmentation backbones. Abbreviations: Comp. = Completion, Acc. = Accuracy, Chamf. = Chamfer distance, Haus. = Hausdorff distance, C.R. = Completion Ratio, F1 = F1-Score, FPS = Frames Per Second. The highlight colors indicate the ranking for each metric: best, second best, and third best.

	ScanNet v2					RIO					Flat						
Method	Comp. [cm](↓)	Acc. [cm](↓)	Chamf. [cm](↓)	Haus. [cm](↓)	C.R. [%](↑)	Comp. [cm](↓)	Acc. [cm](↓)	Chamf. [cm](↓)	Haus. [cm](↓)	C.R. [%](↑)	Comp. [cm](↓)	Acc. [cm](↓)	Chamf. [cm](↓)	Haus. [cm](↓)	C.R. [%](↑)	F1 (↑)	FPS (↑)
GTS	1.38	2.80	9.46	18.53	82.5	1.23	1.91	10.25	21.43	77.09	1.27	0.66	3.05	5.60	71.30	-	-
MaskDINO	1.81	4.40	16.76	35.65	81.21	1.28	1.97	9.16	18.55	76.63	1.26	0.68	3.08	5.58	71.69	76.01	5.95
OpenSeeD	1.64	3.64	13.8	28.95	80.95	1.15	1.89	8.91	18.50	74.39	1.18	0.653	2.71	4.78	66.5	76.43	1.43
OMG-Seg	1.54	4.67	16.46	36.00	76.01	0.92	2.95	9.86	21.60	34.54	1.14	0.67	2.62	4.53	68.03	77.99	1.33
UPPM (Ours)	1.78	3.30	12.64	25.53	81.42	1.19	1.55	5.65	10.05	74.14	1.1	0.61	2.64	4.75	70.76	79.54	2.27

Overall Performance Analysis For comprehensive analysis across diverse scenarios, we present aggregated evaluation results for all datasets using a weighted average approach. As shown in Figure 7, UPPM shows robust performance across multiple metrics including Completion Ratio, F1-score, Chamfer distance, Hausdorff distance, and Accuracy. The Chamfer distance metric is more reliable and less sensitive to outliers than individual Accuracy and Hausdorff distance components, clearly demonstrating that UPPM outperforms other closed-set [5] and open-vocabulary [9,10] approaches in overall performance.

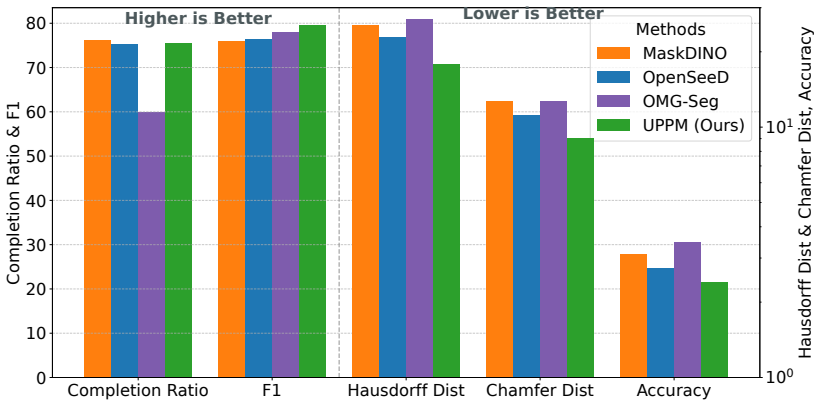

Figure 7. Aggregated performance across ScanNet v2, RIO, and Flat. Completion Ratio and F1-score (left axis, higher is better) and Chamfer distance, Hausdorff distance, and Accuracy (right axis, lower is better) indicate that UPPM offers a favorable trade-off between geometric fidelity and panoptic quality compared with alternative segmentation backbones.

4.2. Map-to-Map Evaluation

UPPM is compared against leading mapping pipelines including Kimera [27], PanMap [24], and DHP [25], representing state-of-the-art approaches in volumetric and semantic mapping. The Chamfer-L1 metric provides robust assessment of overall reconstruction quality by measuring geometric fidelity with reduced sensitivity to outliers compared to L2 norm.

UPPM achieves remarkable performance across all geometric metrics (Table 2), with exceptional accuracy (0.61 cm), completeness (1.10 cm), and Chamfer-L1 distance (1.71 cm). The enhanced accuracy is enabled by the submap-based architecture that supports localized optimization, combined with semantic–geometric integration that improves object-level reconstructions. The better completeness (1.10 cm vs $6.48 - 7.63\,cm$ from competitors) demonstrates UPPM's ability to capture more complete scene geometry through its multi-resolution Multi-TSDF approach, while the outstanding Chamfer-L1 distance (1.71 cm vs $3.62 - 4.24\,cm$) reflects the balanced optimization between accuracy and completeness. These results demonstrate UPPM's effectiveness in producing geometrically accurate maps while maintaining the benefits of dynamic semantic labeling.

 doi:10.20944/preprints202512.0137.v1

Table 2. Geometric reconstruction quality on the Flat dataset. UPPM is compared against state-of-the-art methods, demonstrating exceptional performance in accuracy, completeness, and Chamfer-L1 distance.

Method	**Acc.** [cm]↓	**Comp.** [cm]↓	**Chamfer-L1** [cm]↓
Kimera	0.76	6.48	3.62
Panmap	0.86	7.63	4.24
DHP	0.73	6.58	3.65
UPPM (Ours)	**0.61**	**1.1**	**1.71**

4.3. Segmentation-to-Segmentation Evaluation

As shown in Table 3, UPPM achieves the highest *Panoptic Quality* (PQ) (0.414), *Recognition Quality* (RQ) (0.475), and *mean Average Precision* (mAP) across multiple IoU thresholds, with competitive *Segmentation Quality* (SQ) (0.845). The better PQ and RQ performance stems from the unified semantics approach that reduces label ambiguity and improves instance recognition, while the competitive SQ reflects the precision of instance masks guided by dynamic descriptors and enhanced duplicate suppression through custom NMS. The highest mAP scores across all IoU thresholds (0.549, 0.521, 0.475) demonstrate UPPM's consistent performance in object detection and semantic retrieval, outperforming MaskDINO (0.546, 0.516, 0.470) and other competitors. These results validate the effectiveness of the dynamic labeling approach in producing high-quality panoptic segmentations that maintain both semantic consistency and spatial accuracy.

Table 3. Panoptic segmentation quality on the Flat dataset. UPPM is evaluated against other methods on key metrics, with highlight colors indicating performance ranking: best, second best, and third best.

Method	**PQ** ↑	**RQ** ↑	**SQ** ↑	**mAP-(0.3)** ↑	**mAP-(0.4)** ↑	**mAP-(0.5)** ↑
MaskDINO	0.406	0.470	0.851	0.546	0.516	0.470
OMG-Seg	0.164	0.200	0.498	0.287	0.244	0.200
Detectron2	0.343	0.432	0.787	0.499	0.473	0.432
UPPM (Ours)	0.414	0.475	0.845	0.549	0.521	0.475

4.4. Ablation Studies

The objective of the ablation studies is to gain a deeper understanding of the contributions made by different components in UPPM.

4.5. Unified Semantics

In this study, we examine the effects of deactivating the unified semantic mechanism, which we call PPM (Promptable Panoptic Mapping), to assess the impact of unified semantics on map reconstruction. As shown in Table 4, UPPM outperforms PPM across all metrics on the Flat dataset, achieving better completion, accuracy, and Chamfer distance, in addition to the added benefit of unified semantics. As shown in Figure 8, PPM lacks dynamically labeled classes, leading to ambiguity when many semantic categories attach to the same object. In comparison, UPPM exhibits consistent behavior by reliably assigning semantic classes to objects while maintaining richness in dynamic labeling.

Table 4. Results of Quantitative Ablation Experiments on Flat Dataset. Abbreviations: Comp. = Completion, Acc. = Accuracy, Chamf. = Chamfer distance.

Method	**Comp. [cm](↓)**	**Acc. [cm](↓)**	**Chamf. [cm](↓)**
UPPM	**1.1**	**0.61**	**2.64**
UPPM with tags	1.2	0.63	2.77
PPM	1.18	0.65	2.79
PPM with tags	1.22	0.66	2.92

 doi:10.20944/preprints202512.0137.v1

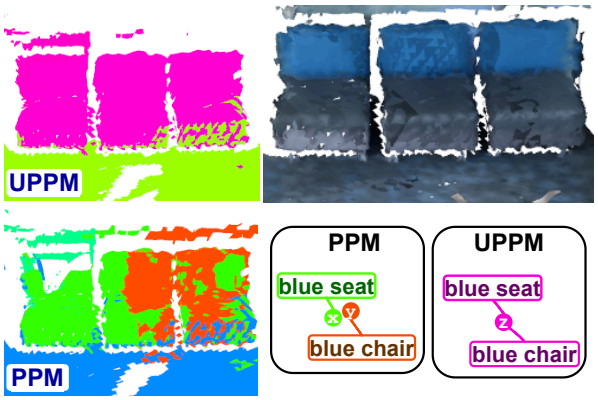

Figure 8. Effect of unified semantics on 3D panoptic segmentation. PPM, which disables unified semantics, assigns multiple ambiguous labels (x, y, z) to the same object, whereas UPPM produces a single consistent semantic class while retaining rich open-vocabulary descriptors.

4.5.1. Non-Maximum Suppression

We compare traditional NMS with our custom NMS on Flat (Tables 5 and 6).

Table 5. Comparison of traditional NMS [52] and our custom NMS on the Flat dataset. Custom NMS demonstrates better performance across all metrics.

Method	Precision [%](↑)	Recall [%](↑)	F1-Score [%](↑)
NMS [52]	91.3	63.64	75.0
Custom NMS (Ours)	**94.6**	**100.0**	**97.22**

Custom NMS improves the completion ratio by 8.27% (62.11%→70.38%) at unchanged Chamfer (2.79 cm), widening reconstructed coverage by removing overlaps and redundancies. The 8.27% gain arises from context-aware suppression that prioritizes caption-derived labels and removes duplicates across semantically equivalent detections. This improvement demonstrates that our custom NMS strategy effectively balances duplicate removal with coverage preservation, achieving better completion without sacrificing geometric accuracy.

Table 6. Quantitative NMS results on the Flat dataset. The custom NMS significantly improves the completion ratio while maintaining the same Chamfer distance.

Method	Chamfer dist. [cm](↓)	Comp. Ratio [<5cm %](↑)
PPM	2.79	**70.38**
PPM w/o NMS	2.79	62.11
Enhancement	0.00	**+8.27 %**

4.5.2. Blurry Frames Filtering

We skip frames flagged as blurry by the caption model (Figure 3). This filtering strategy has negligible effect on Flat, small impact on ScanNet v2 (0.84% flagged), and a larger effect on RIO (3% flagged; manual inspection suggests $>$ 21% affected). On RIO, UPPM reduces Chamfer distance by 16.675% and increases completion ratio by 6.47%. These gains arise because filtering out low-quality frames reduces geometric inconsistencies and error accumulation during fusion.

4.5.3. Identified Tags

As shown in Figure 3, we offer the option to use the identified tags in the pipeline. For the Flat dataset, as shown in Table 4, integrating the identified tags to both PPM and UPPM results in marginal improvement in completion ratio and all other metrics in the case of UPPM. This demonstrates that semantic information adds to more comprehensive scene knowledge.

 doi:10.20944/preprints202512.0137.v1

5. Conclusions

UPPM integrates open-vocabulary vision-language perception with panoptic 3D mapping through a simple dynamic descriptor design. Each object's dynamic descriptor collects rich open-vocabulary labels, maps them to a unified category and size prior, and fuses the result into a multi-resolution multi-TSDF map. This aggregation reconciles the long-tailed semantics of modern foundation models with the consistency requirements of panoptic mapping: it preserves open-vocabulary expressiveness while preventing synonyms from fracturing instances across frames. Experiments on ScanNet v2, RIO and Flat show that UPPM delivers the best reconstruction accuracy and panoptic segmentation quality. It outperforms strong closed-set and open-set baselines, and enables downstream language-conditioned tasks such as object retrieval. Ablation studies attribute these gains to the unified semantics and dynamic descriptors, with custom non-maximum suppression and blurry-frame filtering further improving completeness without sacrificing accuracy.

Apart from the numerous advantages for the proposed method, UPPM inherits limitations from its vision-language backbones. The quality of the captions and detections bounds the accuracy of the dynamic descriptors, and residual semantic drift can persist in highly cluttered or long-tailed settings. The system also incurs non-trivial computational overhead from high-capacity language models and high-resolution map updates. Future work should explore tighter coupling between semantics and geometry, lighter-weight labeling models for real-time deployment, improved temporal reasoning for long-term and multi-session mapping, and extensions to dynamic or outdoor environments.

References

1. McCormac, J.; Handa, A.; Davison, A.; Leutenegger, S. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In Proceedings of the 2017 IEEE International Conference on Robotics and automation (ICRA). IEEE, 2017, pp. 4628–4635.
2. Salas-Moreno, R.F.; Newcombe, R.A.; Strasdat, H.; Kelly, P.H.; Davison, A.J. Slam++: Simultaneous localisation and mapping at the level of objects. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 1352–1359.
3. Narita, G.; Seno, T.; Ishikawa, T.; Kaji, Y. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019, pp. 4205–4212.
4. Sun, Q.; Xu, Z.; Li, Y.; Zhang, Y.; Ye, F. DKB-SLAM: Dynamic RGB-D Visual SLAM with Efficient Keyframe Selection and Local Bundle Adjustment. *Robotics* **2025**, *14*, 134.
5. Li, F.; Zhang, H.; Xu, H.; Liu, S.; Zhang, L.; Ni, L.M.; Shum, H.Y. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3041–3050.
6. Zong, Z.; Song, G.; Liu, Y. Detrs with collaborative hybrid assignments training. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6748–6758.
7. Zhu, Y.; Xiao, N. Simple Scalable Multimodal Semantic Segmentation Model. *Sensors* **2024**, *24*.
8. Zhou, F.; Shi, H. HyPiDecoder: Hybrid Pixel Decoder for Efficient Segmentation and Detection. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 22100–22109.
9. Zhang, H.; Li, F.; Zou, X.; Liu, S.; Li, C.; Yang, J.; Zhang, L. A simple framework for open-vocabulary segmentation and detection. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 1020–1031.
10. Li, X.; Yuan, H.; Li, W.; Ding, H.; Wu, S.; Zhang, W.; Li, Y.; Chen, K.; Loy, C.C. OMG-Seg: Is one model good enough for all segmentation? In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27948–27959.
11. Ghiasi, G.; Gu, X.; Cui, Y.; Hartwig, A.; He, Z.; Khademi, M.; Le, Q.V.; Lin, T.Y.; Susskind, J.; et al. Open-vocabulary Image Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20645–20655.
12. Liu, M.; Ji, J.; Zheng, Z.; Han, K.; Shan, X.; Wang, X.; Wang, L.; Wang, D.; Gao, P.; Zhang, H.; et al. OpenSeg: Panoramic Open-Vocabulary Scene Parsing with Vision-Language Models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 17006–17016.

 doi:10.20944/preprints202512.0137.v1

13. Li, J.; Ding, H.; Tian, X.; Li, X.; Long, X.; Yuan, Z.; Li, J. OVSeg: Label-Free Open-Vocabulary Training for Text-to-Region Recitation. In Proceedings of the Advances in Neural Information Processing Systems, 2023.
14. Chen, H.; Blomqvist, K.; Milano, F.; Siegwart, R. Panoptic Vision-Language Feature Fields. *IEEE Robotics and Automation Letters* **2024**, *9*, 2144–2151.
15. OpenVox Authors. OpenVox: Real-time Instance-level Open-Vocabulary Probabilistic Voxel Representation. *arXiv preprint* **2024**. Available online: arXiv (accessed on 15 May 2024).
16. Jatavallabhula, K.M.; Kuwajerwala, A.; Gu, Q.; Omama, M.; Chen, T.; Maalouf, A.; Li, S.; Iyer, G.S.; Keetha, N.V.; Tewari, A.; et al. ConceptFusion: Open-set Multimodal 3D Mapping. In Proceedings of the ICRA2023 Workshop on Pretraining for Robotics (PT4R), 2023.
17. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 8748–8763.
18. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660.
19. Gu, Q.; Kuwajerwala, A.; Morin, S.; Jatavallabhula, K.M.; Sen, B.; Agarwal, A.; Rivera, C.; Paul, W.; Ellis, K.; Chellappa, R.; et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 5021–5028.
20. Clio Authors. Clio: Language-Guided Object Discovery in 3D Scenes. *arXiv preprint* **2024**. Available online: arXiv (accessed on 15 May 2024).
21. Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. Openscene: 3d scene understanding with open vocabularies. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 815–824.
22. Huang, X.; Zhang, Y.; Ma, J.; Tian, W.; Feng, R.; Zhang, Y.; Li, Y.; Guo, Y.; Zhang, L. Tag2Text: Guiding Vision-Language Model via Image Tagging. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
23. Honnibal, M. A Good Part-of-Speech Tagger in about 200 Lines of Python. https://explosion.ai/blog/part-of-speech-pos-tagger-in-python, 2013.
24. Schmid, L.; Delmerico, J.; Schönberger, J.L.; Nieto, J.; Pollefeys, M.; Siegwart, R.; Cadena, C. Panoptic multi-tsdfs: a flexible representation for online multi-resolution volumetric mapping and long-term dynamic scene consistency. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022, pp. 8018–8024.
25. Hu, T.; Jiao, J.; Xu, Y.; Liu, H.; Wang, S.; Liu, M. Dhp-mapping: A dense panoptic mapping system with hierarchical world representation and label optimization techniques. In Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024, pp. 1101–1107.
26. McCormac, J.; Clark, R.; Bloesch, M.; Davison, A.; Leutenegger, S. Fusion++: Volumetric object-level slam. In Proceedings of the 2018 international conference on 3D vision (3DV). IEEE, 2018, pp. 32–41.
27. Rosinol, A.; Abate, M.; Chang, Y.; Carlone, L. Kimera: an open-source library for real-time metric-semantic localization and mapping. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020, pp. 1689–1696.
28. Mahmoud, A.; Atia, M. Improved visual SLAM using semantic segmentation and layout estimation. *Robotics* **2022**, *11*, 91.
29. Laina, S.B.; Boche, S.; Papatheodorou, S.; Schaefer, S.; Jung, J.; Leutenegger, S. FindAnything: Open-Vocabulary and Object-Centric Mapping for Robot Exploration in Any Environment. *arXiv preprint arXiv:2504.08603* **2025**.
30. Martins, T.B.; Oswald, M.R.; Civera, J. Open-vocabulary online semantic mapping for SLAM. *IEEE Robotics and Automation Letters* **2025**.
31. Patel, N.; Krishnamurthy, P.; Khorrami, F. RAZER: Robust Accelerated Zero-Shot 3D Open-Vocabulary Panoptic Reconstruction with Spatio-Temporal Aggregation. *arXiv preprint arXiv:2505.15373* **2025**.
32. Yang, D.; Wang, X.; Gao, Y.; Liu, S.; Ren, B.; Yue, Y.; Yang, Y. OpenGS-Fusion: Open-Vocabulary Dense Mapping with Hybrid 3D Gaussian Splatting for Refined Object-Level Understanding. *arXiv preprint arXiv:2508.01150* **2025**.

 doi:10.20944/preprints202512.0137.v1

33. Jiang, J.; Zhu, Y.; Wu, Z.; Song, J. DualMap: Online Open-Vocabulary Semantic Mapping for Natural Language Navigation in Dynamic Changing Scenes. *IEEE Robotics and Automation Letters* **2025**, *10*, 12612–12619. https://doi.org/10.1109/LRA.2025.3621942.
34. Gu, Q.; Ye, Z.; Yu, J.; Tang, J.; Yi, T.; Dong, Y.; Wang, J.; Cui, J.; Chen, X.; Wang, Y. MR-COGraphs: Communication-Efficient Multi-Robot Open-Vocabulary Mapping System via 3D Scene Graphs. *IEEE Robotics and Automation Letters* **2025**, *10*, 5713–5720. https://doi.org/10.1109/LRA.2025.3561569.
35. Wei, W.; Ülger, O.; Nejadasl, F.K.; Gevers, T.; Oswald, M.R. 3D-AVS: LiDAR-based 3D Auto-Vocabulary Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2025, pp. 8910–8920.
36. Zhang, C.; Delitzas, A.; Wang, F.; Zhang, R.; Ji, X.; Pollefeys, M.; Engelmann, F. Open-vocabulary functional 3d scene graphs for real-world indoor spaces. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 19401–19413.
37. LangSplat Authors. LangSplat: Language-Conditioned 3D Gaussian Splatting. *arXiv preprint* **2024**. Available online: arXiv (accessed on 15 May 2024).
38. Bayesian-Fields Authors. Bayesian-Fields: Probabilistic 3D Scene Understanding with Open-Vocabulary Queries. *arXiv preprint* **2023**. Available online: arXiv (accessed on 15 May 2024).
39. OpenGaussian Authors. OpenGaussian: Open-Vocabulary 3D Gaussian Splatting. *arXiv preprint* **2024**. Available online: arXiv (accessed on 15 May 2024).
40. Werby, A.; Huang, C.; Büchner, M.; Valada, A.; Burgard, W. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In Proceedings of the First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024, 2024.
41. Awais, M.; Naseer, M.; Khan, S.; Anwer, R.M.; Cholakkal, H.; Shah, M.; Yang, M.H.; Khan, F.S. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.
42. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 4015–4026.
43. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* **2023**.
44. Lu, S.; Chang, H.; Jing, E.P.; Boularias, A.; Bekris, K. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In Proceedings of the Conference on Robot Learning. PMLR, 2023, pp. 1610–1620.
45. Takmaz, A.; Fedele, E.; Sumner, R.W.; Pollefeys, M.; Tombari, F.; Engelmann, F. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
46. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In Proceedings of the European conference on computer vision. Springer, 2024, pp. 38–55.
47. Song, K.; Tan, X.; Qin, T.; Lu, J.; Liu, T.Y. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems* **2020**, *33*, 16857–16867.
48. Zhang, Y.; Huang, X.; Ma, J.; Li, Z.; Luo, Z.; Xie, Y.; Qin, Y.; Luo, T.; Li, Y.; Liu, S.; et al. Recognize anything: A strong image tagging model. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1724–1732.
49. Caesar, H.; Uijlings, J.; Ferrari, V. Coco-stuff: Thing and stuff classes in context. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1209–1218.
50. Fellbaum, C. WordNet. In *Theory and applications of ontology: computer applications*; Springer, 2010; pp. 231–243.
51. Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data* **2019**, *7*, 535–547.
52. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
53. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5828–5839.

 doi:10.20944/preprints202512.0137.v1

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