PARTIMAP: ONLINE OPEN-VOCABULARY 3D PANOPTIC MAPPING AS PARTITION INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Online open-vocabulary 3D panoptic mapping is a fundamental task in scene understanding. The central difficulty is fusing partial and noisy 2D masks from many views into consistent 3D instances. Existing online systems fuse with hand-designed rules that lack an explicit objective, which discard uncertainty of 2D observations, lead to potentially sub-optimal merging, and require special treatment if estimated poses change. We present PartIMap, which formulates mask fusion as probabilistic inference. PartIMap maintains a probabilistic voxel map in which semantics are von Mises-Fisher distributions on the descriptor sphere, kept per instance and per voxel, and instance geometry is a per-voxel Beta occupancy state. Deciding which masks belong to the same object is cast as maximum a posteriori inference of a partition on a mask graph. This partition model retains probabilistic data association as its first-order relation while capturing higher-order relations such as transitivity, and yields soft membership weights from the same posterior. Every fused quantity is an additive sum over its constituent observations, and the map can be updated effectively and efficiently by re-merging when a live SLAM system closes a loop. PartIMap is zero-shot and requires no supervised training. On ScanNetV2 and Replica it outperforms both online and offline methods, improving the strongest online baseline by 3.2 AP and 3.9 PRQ (T) on ScanNetV2 at comparable speed, degrades the least when poses come from live SLAM, and exposes per-voxel uncertainty. Codes will be released upon publication.

1 INTRODUCTION

Online open-vocabulary 3D scene understanding is crucial for various embodied AI tasks (Mousavian et al., 2019; James et al., 2020; Matsuki et al., 2024). This has become practical because 2D foundation models (Kirillov et al., 2023; Radford et al., 2021; Li et al., 2022) provide class-agnostic masks and language-aligned descriptors for any image, which bypasses the scarcity of annotated 3D data. The challenge lies in fusion: each frame observes only a fragment of the scene, 2D masks are inconsistent across views and frequently over-segmented. The system must decide, as observations arrive, which masks describe the same object and how their evidence should be combined.

Existing online systems (Tang et al., 2025; Zhai et al., 2026; Martins et al., 2025; Deng et al., 2026) answer this question with hand-designed hard criteria, such as similarity thresholds, multiple rules, or bipartite assignment. Such hand-crafted merging methods have three limitations, as illustrated in Fig. 1. First, they discard the uncertainty inherent in partial 2D observations, merging ambiguous observations with the same confidence as good ones. Second, the rules are heuristics whose thresholds and structure approximate an unstated and often vague objective, so that the resulting grouping is sub-optimal. Third, when operating on a live backend and loop-closure updates camera poses, they either freeze the map or rely on special rules to correct drifted segmentation.

Probabilistic data association (Bar-Shalom et al., 2009; Bowman et al., 2017) is a principled tool for this problem. It replaces hard assignment with association probabilities, so that uncertain observations contributes less than confident ones to a merging decision. Its literature, however, concerns only closed-set vocabularies. More importantly, data association is mainly applied to mapping sparse landmarks, since it models the *matching* between observation and existing objects. In contrast, we argue that to merge dense segmentations, the latent structure is a *partition* on the mask graph.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this paper we present PartIMap, an online open-vocabulary 3D panoptic mapping model that treats fusion as inference in a probabilistic model, illustrated in Fig. 2. PartIMap maintains a probabilistic voxel map: semantics are von Mises–Fisher distributions kept both per instance and per voxel, and instance geometry is a per-voxel Beta occupancy state. Both are defined by additive sufficient statistics, so that fusing an observation is a weighted sum. Mask merging is cast as maximum a posteriori inference of a partition on a mask graph, whose posterior retains pairwise probabilistic association as its first-order relation, and models transitivity and evidence accumulation without additional rules. The same posterior yields a soft membership weight for every observation, so that a marginal merge affects the fused object only in proportion to its evidence. Finally, because every fused state is an exact sum over its constituent observations, PartIMap runs on a live ORB-SLAM3 back-end (Campos et al., 2021) and merges drifted segments when a loop closure corrects the trajectory, by re-running the same partition inference on the corrected geometry.

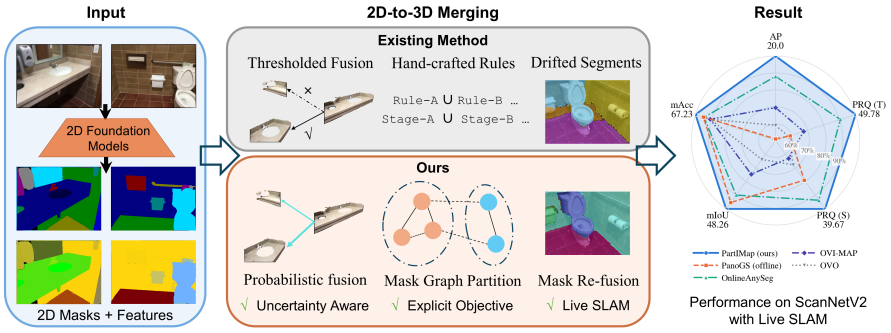

Figure 1: Online systems use hand-designed 2D mask fusion, which discards the uncertainty of partial observations, has sub-optimal solutions, and needs special treatment when camera pose changes. PartIMap infers in a single probabilistic model: uncertainty modeled by a membership weight, solution found via mask graph partition, and same model used for loop-closure re-fusion.

Our contributions are as follows.

- A probabilistic voxel representation for open-vocabulary panoptic mapping, combining von Mises–Fisher semantics at the instance and voxel level with Beta occupancy, whose additive statistics support merging and re-fusion, and which exposes per-voxel semantic and spatial uncertainty at no additional cost.
- A partition model for mask merging that generalizes probabilistic data association. Its estimator and membership weights are derived from the same posterior.
- Integration with live SLAM through a per-observation provenance invariant, which facilitates loop-closure re-fusion for drifted segments, and requires no special machinery.
- State-of-the-art results on ScanNetV2 and Replica among both online and offline methods, with gains of 3.2 AP and 3.9 PRQ (T) over the strongest online baseline on ScanNetV2 at comparable speed, the smallest degradation under live SLAM poses, and consistent improvements across 2D backbones, all without training.

2 Related Work

3D instance and semantic segmentation from foundation models. The scarcity of annotated 3D data has motivated transferring 2D foundation models such as SAM (Kirillov et al., 2023) and CLIP (Radford et al., 2021) to 3D scene understanding. Offline methods distill 2D features and lift them to 3D, on a fully reconstructed scene (Peng et al., 2023; Takmaz et al., 2023; Yang et al., 2023; Yan et al., 2024), or embed them into radiance fields or 3D Gaussians (Kerr et al., 2023; Qin et al., 2024; Ye et al., 2024; Zhai et al., 2025). In contrast, online methods address instance segmentation from streaming RGB-D. Early models fuse per-frame closed-vocabulary predictions into an incrementally growing point cloud (Liu et al., 2022; Xu et al., 2024). EmbodiedSAM (Xu et al., 2025) lifts SAM masks to superpoint queries, and AutoSeg3D (Wang et al., 2025a) recasts the problem as instance tracking. Both are trained on a fixed label set, and adapting them to open-vocabulary would

require re-design and re-training. A parallel line of work (Fan et al., 2024; Sun et al., 2026; Zhu et al., 2026; Li et al., 2026; Lee et al., 2026) lifts scene understanding into feed-forward reconstruction backbones (Wang et al., 2024; 2025b). They are bound to the input regime of their underlying geometry model, *i.e.*, a small set of selected images. We focus on the complementary setting of building a 3D panoptic map online and incrementally, without task-specific training.

Online 3D panoptic segmentation. Online semantic mapping originated in the closed-set setting (McCormac et al., 2017; Narita et al., 2019; Mccormac et al., 2018; Grinvald et al., 2019). Open-vocabulary variants (Jatavallabhula et al., 2023; Yamazaki et al., 2024; Gu et al., 2024) replace class labels with vision-language features. Recent systems further enable instance-level panoptic mapping. OnlineAnySeg (Tang et al., 2025) organizes 2D masks in a hashed voxel volume and merges them when a thresholded similarity is exceeded, or when enough third masks support the merge. OnlinePG (Zhai et al., 2026) clusters 3D Gaussian segments within a sliding window and fuses them by bipartite matching. OVO (Martins et al., 2025) integrates open-vocabulary mapping into a SLAM system with a learned merging network, and loop closure re-merging uses fixed similarity thresholds. OVI-MAP (Deng et al., 2026) decouples a class-agnostic TSDF instance map from semantic inference, which is performed on a small set of selected views. Despite their different representations, all of these systems decide association with hard rules, namely binary overlap gates, thresholded similarities, or one-to-one assignments, whose criteria are designed by hand. Such decisions discard the uncertainty inherent in partial 2D observations, and a merge made on weak evidence can potentially contaminate the fusion. PartIMap instead maintains a probabilistic representation of both semantics and geometry, and infers mask grouping under a single posterior.

Revisiting data association. Data association, *i.e.*, deciding which measurements originate from the same object, is a classical problem in target tracking where data correspondence is uncertain (Reid, 1979; Mahler, 2003; Bar-Shalom et al., 2009). These formulations are widely used in multi-object tracking (Luo et al., 2021). In 3D scene understanding, Bowman et al. (2017) treat the association of landmarks as a latent variable of semantic SLAM and optimize it jointly with pose, EM-Fusion (Strecke & Stuckler, 2019) infers pixel-to-object association likelihoods for dynamic object-level SLAM, ConvBKI (Wilson et al., 2024) fuses semantic predictions into a voxel map with conjugate Dirichlet updates that quantify uncertainty, and PUF (Yang et al., 2026) applies probabilistic node association with Dirichlet evidence accumulation to online scene graph generation. These methods operate on a closed-set vocabulary, on a sparse set of landmark objects, or on a per-observation matching between the current frame and the map. Our work extends probabilistic data association to dense, open-vocabulary panoptic mapping from foundation model outputs, and generalizes the matching formulation to partition inference on a mask graph.

3 METHOD

3.1 OVERVIEW

Problem setting. The input is a stream of RGB-D keyframes whose camera poses are provided by either ground truth or a live SLAM system. Our goal is to maintain, at every time step, a voxelized truncated signed distance function (TSDF) reconstruction of the observed scene, together with a set of open-vocabulary 3D instances. For every keyframe, a class-agnostic 2D segmenter proposes a set of masks, and a dense vision-language model provides one d-dimensional descriptor per mask by masked mean pooling. Each mask is back-projected through the depth map into the voxel grid, which yields an observation o defined by a voxel set V_o, a semantic descriptor $\boldsymbol{x}_o \in \mathbb{S}^{d-1}$, and the index of its source keyframe. Each 3D instance k is assembled from its associated 2D mask observations, with its semantic and spatial descriptor, as well as the instance-mask association weights defined in Sec. 3.2. Observations and instances are held in a mask bank built on a hashed voxel volume, from which a mask graph is formed at every merge step; Appendix A.1 defines these structures, and Appendix A.2 gives the complete online procedure.

Pipeline. The method comprises three stages. **(i) Representation** (Sec. 3.2). Every observation is fused into a probabilistic voxel map. Semantics are represented by von Mises-Fisher (vMF) distributions on the descriptor sphere, and instance geometry is represented by a per-voxel Beta occupancy state. **(ii) Merging** (Sec. 3.3). Deciding which observations belong to the same object is

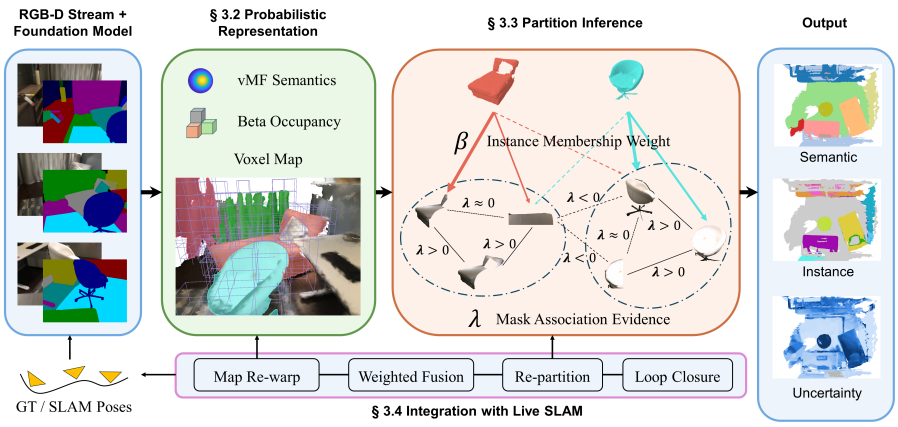

Figure 2: Overview of PartIMap. Each 2D mask is projected to the voxel grid as an observation and fused into a probabilistic voxel map (§3.2), whose states are additive in the per-observation weights β_o. At a fixed step the current instances and the newly inserted masks form a mask graph, and merging is MAP inference of the graph partition (§3.3), which yields the clusters and the membership weights. Loop closure is handled by re-deriving the map and re-running one partition step (§3.4).

cast as approximate maximum a posteriori (MAP) inference of a partition of the current mask graph. **(iii) Live SLAM** (Sec. 3.4). Camera poses are obtained from a live ORB-SLAM3 back-end (Campos et al., 2021). At every loop closure, the map is re-integrated under the corrected trajectory and the partition inference is re-run on the corrected geometry.

3.2 Probabilistic Voxel Representation

vMF semantics. We model the descriptors of the observations belonging to one object as samples from a von Mises–Fisher distribution on $\mathbb{S}^{d-1}$, $\mathrm{vMF}(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) \propto \exp(\kappa\, \boldsymbol{\mu}^\top \boldsymbol{x})$, with mean direction $\boldsymbol{\mu}$ and concentration κ. Its sufficient statistics are additive, so an instance k is fully described by a weighted resultant vector and a total weight,

$$\boldsymbol{R}_k = \sum_{o \in k} \beta_o\, \boldsymbol{x}_o, \qquad W_k = \sum_{o \in k} \beta_o, \tag{1}$$

where $\beta_o \in (0, 1]$ is the membership weight assigned to observation o when it was merged (Sec. 3.3). The maximum-likelihood mean direction and mean resultant length follow as $\hat{\boldsymbol{\mu}}_k = \boldsymbol{R}_k / \|\boldsymbol{R}_k\|$ and $\bar{r}_k = \|\boldsymbol{R}_k\| / W_k$ (Banerjee et al., 2005). Merging instances amounts to summing their statistics. The instance descriptor for association and open-vocabulary classification is $\hat{\boldsymbol{\mu}}_k$.

Per-voxel language grid. An instance-level descriptor assigns a single label to every voxel covered by the instance, which is fragile where instance grouping fragments. We therefore maintain the same vMF accumulator one level down, as a sparse grid over occupied voxels: each voxel v stores a resultant $\boldsymbol{G}_v = \sum_{o:\, v \in V_o} \boldsymbol{x}_o$ and an observation count $N_v = |\{o : v \in V_o\}|$. The grid is populated directly from raw mask observations at insertion time and is therefore decoupled from merge decisions. It is stored as a sparse index over allocated voxels with chunk-grown half-precision rows, so that only observed voxels consume memory. Because the grid is written independently of grouping, an incorrect merge does not contaminate it.

Beta occupancy. Prior online systems represent instance geometry as a binary voxel set, so that a voxel claimed by a marginal merge has equal evidence as a voxel observed from multiple views. We replace the binary set with per-voxel Beta evidence over the support of the instance,

$$a_v^k = \sum_{o \in k} \beta_o\, \mathbf{1}[v \in V_o], \qquad \mathbb{E}[\mathrm{occ}_v^k] = \frac{a_v^k}{a_v^k + b_0}, \tag{2}$$

where b_0 is a prior pseudo-count; $b_0 \to 0$ recovers binary membership. Eq. (2) and Eq. (1) share the same weights β_o and the same additivity, so they are maintained by identical machinery. The expected occupancy also provides a soft spatial cue for mask merging (Sec. 3.3).

Decoding. For semantic output, we classify per-voxel directions against the text embeddings of the vocabulary. A raw per-voxel argmax assigns spurious rare labels to well-observed voxels whose local mean deviates slightly from the object consensus. We therefore regularize each voxel toward the instance $k(v)$ that covers it, using $\boldsymbol{\mu}_v^{\text{final}} = \text{norm}(\boldsymbol{G}_v + \boldsymbol{R}_{k(v)})$. Per-voxel occupancy $\mathbb{E}[\text{occ}_v]$ and consistency $\bar{r}_v = \|\boldsymbol{G}_v\|/N_v$ are exported alongside the labels.

3.3 Mask merging as Partition inference

Partition model. At every merge step, we represent the current masks as a graph whose n nodes are the existing instances, including the new observations inserted since the last step as singleton instances, and whose edge set $\mathcal{E}$ carries pairwise evidence about whether two nodes belong to the same object. Appendix A.1 describes how the graph is built and maintained.

Let $z_a \in \{1, \dots, K\}$ denote the instance identity of node a, with the number of instances K unknown, and let $X_{ab} = \mathbf{1}[z_a = z_b]$ be the co-membership relation between nodes a and b. X encodes a partition of the nodes, and mask merging becomes a correlation clustering problem (Bansal et al., 2004). We model the pairwise evidence y_{ab} on the edges as conditionally independent given the partition, and define the mask association evidence λ_{ab} on each edge as the log-likelihood ratio

$$p(Y \mid X) = \prod_{(a,b)\in\mathcal{E}} p(y_{ab} \mid X_{ab}), \qquad \lambda_{ab} := \log \frac{p(y_{ab} \mid X_{ab} = 1)}{p(y_{ab} \mid X_{ab} = 0)}. \tag{3}$$

As a prior, we assign a cost $\theta > 0$ to every co-membership asserted on an edge of the graph, $p(X) \propto \exp\big(-\theta \sum_{(a,b)\in\mathcal{E}} X_{ab}\big)$, which is a Gibbs distribution over the finite set of partitions of the nodes. The log-posterior is then

$$\log p(X \mid Y) = \sum_{(a,b)\in\mathcal{E}} \big(\lambda_{ab} - \theta\big)\, X_{ab} \;+\; \text{const}, \tag{4}$$

a correlation clustering objective with edge weights $\lambda_{ab} - \theta$; the derivation is given in Appendix B. The prior is applied to the edges, so the criterion for merging two clusters does not depend on their sizes (Eq. (7)). Eq. (4) models transitivity: if $\lambda_{ac} - \theta$ and $\lambda_{bc} - \theta$ are positive, then a, b, c automatically share a cluster, unless $\lambda_{ab} << \theta$. Conversely, conditioning Eq. (4) on a single node recovers the probabilistic data association weights, as shown in Appendix C.

Pairwise evidence. For each pair (a, b) we compute an additive association score from four cues,

$$s_{ab} = s_{\text{se}}(a, b) + s_{\text{occ}}(a, b) + s_{\text{geo}}(a, b) + s_{\text{sp}}(a, b). \tag{5}$$

each cue in $[0, 1]$. The semantic cue s_{se} is the cosine similarity between the vMF means, $\hat{\boldsymbol{\mu}}_a^\top \hat{\boldsymbol{\mu}}_b$ (see Appendix D for its calibration). The occupancy cue is a soft containment ratio weighted by the expected occupancy of Eq. (2), $s_{\text{occ}}(a, b) = |V_b|^{-1} \sum_{v\in V_a\cap V_b} \mathbb{E}[\text{occ}_v^a]$, so that an ambiguous merge drives $\mathbb{E}[\text{occ}]$ toward $\frac{1}{2}$.

Following Tang et al. (2025), the geometric cue s_{geo} is the cosine similarity of FCGF features (Choy et al., 2019), and the spatial cue $s_{\text{sp}} = \frac{1}{2}(c_{ab} + c_{ba})$ symmetrizes the visibility-gated containment ratio c_{ab} (Appendix A.1, Eq. (9)). We model the log-likelihood ratio of Eq. (3) as

$$\lambda_{ab} = T\,(s_{ab} - s_{\text{stay}}), \qquad (a, b) \in \mathcal{E}, \tag{6}$$

where s_{stay} is the indifference level and T the calibration slope. The edge set contains only co-visible pairs, so two masks never observed together contribute 0, since their pairwise cues carry no information about whether they belong to the same object.

Estimator. Exact MAP inference of Eq. (4) is NP-hard (Bansal et al., 2004). In our setting, co-visibility splits the nodes into small connected components, so we apply greedy agglomeration on the objective. Starting from singletons, we repeatedly merge the pair of clusters (P, Q) with the largest gain as long as $\Delta > 0$:

$$\Delta(P, Q) = \sum_{(a,b)\in\mathcal{E}_{PQ}} \big(\lambda_{ab} - \theta\big) = |\mathcal{E}_{PQ}|\,\big(\bar{\lambda}_{PQ} - \theta\big), \qquad \mathcal{E}_{PQ} = \mathcal{E} \cap (P \times Q), \tag{7}$$

where $\bar{\lambda}_{PQ}$ is the mean evidence over the edges between the two clusters. $\Delta(P, Q)$ is the exact change of Eq. (4) caused by the merge. A merge is therefore accepted if and only if $\bar{\lambda}_{PQ} > \theta$.

Membership weights. Once a cluster has been decided, from Eq. (4) each member a receives its posterior probability that a belongs to cluster $\mathcal{C}(a)$, with which its statistics enter Eqs. (1) and (2),

$$\beta_a = \sigma\Big(\sum_{b\in\mathcal{C}(a)\cap\mathcal{N}(a)} \big(\lambda_{ab} - \theta\big)\Big), \tag{8}$$

where $\mathcal{N}(a)$ denotes the neighbors of a in the mask graph, σ the logistic function, and $\beta_a = 1$ for singletons and for the cluster seed. The derivation is given in Appendix B.

3.4 Live SLAM and loop-closure re-fusion

Existing online systems that operate on a live SLAM system either leave the map unchanged at loop closure or re-warp the stored points and repeat a thresholded merge (Martins et al., 2025). Our probabilistic voxel map (Sec. 3.2) supports live update which efficiently re-uses our mask merging at loop closure, with instance segmentation corrected by mask merging (Sec. 3.3).

Per-observation provenance. For every accepted observation o, we store its raw descriptor $\boldsymbol{x}_o$, its source keyframe, its voxel set, and its effective weight β_o within the instance that currently holds it. When instances are merged with the member weights of Eq. (8), the weight is folded multiplicatively into the β_o of the constituent observations, so that Eqs. (1,2) hold at all times. Every index structure of the map (Appendix A.1) is a view over this append-only bank of raw observations, and only the observation-instance ID mapping is modified at merging. Therefore, any instance can be rebuilt in corrected geometry from its constituent observations, which is the basis of the SLAM extension.

Loop-closure re-fusion. On a loop-closure event with corrected keyframe poses, the handler (Algorithm 2) propagates the sparse keyframe corrections to all integrated frames, re-integrates the TSDF under the corrected poses, re-warps the voxel sets of the raw observation bank, and rebuilds index structure. Instance statistics and the occupancy state are corrected from Eqs. (1,2) and the language grid is re-scattered from the stored descriptors. The full containment matrix is then recomputed from the corrected geometry, and one ordinary mask graph partition step (Sec. 3.3) is run over all instances. No loop-closure-specific merge machinery is required. Details of pose tracking and anchoring are given in Appendix D.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on two widely used indoor RGB-D datasets, ScanNetV2 (Dai et al., 2017) and Replica (Straub et al., 2019). We use the same 10 sequences as (Zhai et al., 2025; 2026) on ScanNetV2, and the commonly adopted 8 scenes on Replica. PartIMap's predictions are projected onto the ground-truth point cloud for evaluation, following Tang et al. (2025) and Zhai et al. (2026).

Evaluation metrics. For instance segmentation, we report class-agnostic average precision (AP), averaged over IoU thresholds from 50% to 95%, following OnlineAnySeg (Tang et al., 2025). For semantic segmentation, we report point-level mean intersection over union (mIoU) and mean accuracy (mAcc). For panoptic segmentation, we report the 3D panoptic reconstruction quality (Dahnert et al., 2021) at thing level, PRQ (T), and stuff level, PRQ (S), following PanoGS (Zhai et al., 2025) and OnlinePG (Zhai et al., 2026). Runtime is reported in frames per second (FPS) and excludes the 2D foundation models, following OnlineAnySeg and OnlinePG. For the live SLAM setting, we additionally report the average relative decline (Decl.) of the five metrics with respect to the ground-truth-pose result of the same method, which measures robustness to pose error.

Implementation details. We use LSeg (Li et al., 2022) as the pixel-aligned vision-language model, following PanoGS and OnlinePG, and CropFormer (Qi et al., 2023) as the class-agnostic 2D segmenter, following OnlineAnySeg and OVI-MAP (Deng et al., 2026). Baselines are reproduced under these same 2D backbones unless explicitly stated. The prior weight is set to $\theta = 3$ (Sec. 4.3). In the live SLAM setting, poses are provided by ORB-SLAM3 (Campos et al., 2021) in RGB-D mode (Appendix D). All methods are evaluated on a single NVIDIA RTX 3090 GPU and an Intel(R) Core(TM) i9-9900K CPU.

4.2 Quantitative Results

Ground-truth poses. Table 1 reports results with ground-truth poses. Our online and zero-shot method achieves the best or second-best result on all metrics. On ScanNetV2, the offline method PanoGS, retains a small advantage in mIoU, whereas our method is substantially more accurate in AP (25.7 vs. 14.1) and in both panoptic metrics. On Replica, OVO, which is not zero-shot, obtains a marginally higher mAcc than our method but over-segments the scene, as its low PRQ (T) reflects, and PanoGS retains the best PRQ (S). In terms of runtime, our method runs at 6.9 FPS, on par with OnlineAnySeg (7.0 FPS) and comparable to simpler fusion pipelines such as ConceptGraphs (8.4 FPS). The partition inference thus runs at comparable speed as hand-designed merging rules.

Table 1: 3D instance, semantic, and panoptic segmentation with ground-truth poses on ScanNetV2 and Replica. *: taken from Zhai et al. (2026). Best results are in **bold**, second best are underlined.

Method	Online	Zero-shot	ScanNetV2					Replica					FPS
			AP	mIoU	mAcc	PRQ(T)	PRQ(S)	AP	mIoU	mAcc	PRQ(T)	PRQ(S)	
OpenScene* (Peng et al., 2023)	×	×	–	47.63	69.74	43.53	40.43	–	49.03	62.89	33.04	11.84	–
Open3DIS (Nguyen et al., 2024)	×	✓	11.6	26.41	57.69	36.09	26.48	8.7	41.78	58.55	17.65	16.46	–
InstanceGaussian* (Li et al., 2025)	×	✓	–	34.14	54.95	39.04	27.41	–	–	–	–	–	–
PanoGS (Zhai et al., 2025)	×	✓	14.1	**50.38**	70.15	33.69	36.52	15.6	54.42	67.38	42.85	**30.52**	–
OVO (Martins et al., 2025)	✓	×	16.7	34.91	62.57	34.61	28.61	16.2	54.21	**69.49**	14.63	13.35	8.0
ConceptGraphs (Gu et al., 2024)	✓	✓	14.3	30.68	56.60	32.48	28.98	12.5	42.12	54.52	16.22	20.68	8.4
OnlineAnySeg (Tang et al., 2025)	✓	✓	22.5	45.33	65.85	47.90	42.36	15.9	54.13	66.86	40.39	22.16	7.0
OnlinePG* (Zhai et al., 2026)	✓	✓	–	48.48	66.01	37.97	41.81	–	47.93	54.94	41.02	12.83	–
OVI-MAP (Deng et al., 2026)	✓	✓	16.0	39.93	64.96	36.32	27.43	14.5	52.98	63.64	21.27	16.71	7.3
PartIMap(ours)	✓	✓	**25.7**	49.06	**70.33**	**51.84**	**44.90**	**16.4**	**56.10**	67.80	**43.50**	24.45	6.9

Table 2: Segmentation with poses estimated by live SLAM on ScanNetV2 and Replica. †: OnlineAnySeg does not support live SLAM; we equip it with our re-fusion mechanism. Decl.: average relative decline (%) of the five metrics with respect to the ground-truth-pose setting of the same method (Table 1). OnlinePG has no open-source code and is omitted.

Method	Input	Zero-shot	ScanNetV2						Replica						FPS
			AP	mIoU	mAcc	PRQ(T)	PRQ(S)	Decl.↓	AP	mIoU	mAcc	PRQ(T)	PRQ(S)	Decl.↓	
Open3DIS (Nguyen et al., 2024)	Replay	✓	9.6	24.17	53.93	34.98	25.20	**8.0**	7.7	39.63	56.47	16.12	15.67	6.7	–
PanoGS (Zhai et al., 2025)	Replay	✓	10.2	45.98	63.94	29.39	31.32	14.4	14.6	53.32	66.82	38.10	**28.67**	5.3	–
OVO (Martins et al., 2025)	Live+LC	×	11.8	30.80	61.09	28.94	26.87	13.2	15.6	54.21	**69.04**	13.95	12.36	3.3	7.8
ConceptGraphs (Gu et al., 2024)	Replay	✓	9.5	28.00	53.61	30.55	23.73	14.3	12.1	40.16	54.26	12.49	20.09	6.8	8.4
OnlineAnySeg† (Tang et al., 2025)	Live+LC	✓	17.6	43.37	63.65	45.02	37.15	9.6	14.8	53.28	66.14	38.56	21.23	3.7	6.9
OVI-MAP (Deng et al., 2026)	Replay	✓	13.9	36.00	61.05	33.62	25.00	9.1	12.7	50.64	59.02	18.80	13.94	10.5	7.3
PartIMap(ours)	Live+LC	✓	**20.0**	**48.26**	**67.23**	**49.78**	**39.67**	8.8	**16.0**	**55.47**	67.40	**41.39**	26.04	**0.5**	6.9

Live SLAM. Table 2 reports results with poses estimated by ORB-SLAM3. Methods marked *Replay* have no live SLAM capability and are run offline on the final loop-closure-corrected poses, which is favorable for them (see Appendix E). OnlineAnySeg does not support live SLAM either; for a fair comparison, we equip it with our re-fusion mechanism of Sec. 3.4 but use its own mask merging rules. Our method remains the most accurate on eight segmentation metrics and is second on the remaining two. Its average decline is the smallest on Replica and the second smallest on ScanNetV2, behind Open3DIS. Note that the relative decline favors methods with low absolute accuracy, since they have less room to degrade. Our method, being also the most accurate, therefore demonstrates its robustness to noisy poses. Finally, the frame rate is unchanged with respect to the ground-truth-pose setting (6.9 FPS), which shows that the loop-closure re-fusion is efficient.

4.3 Ablation Study

Backbone sensitivity. To verify that the improvements are not tied to a particular 2D backbone, we replace CropFormer and LSeg with SAM (Kirillov et al., 2023) masks and CLIP (Radford et al., 2021) descriptors for both OnlineAnySeg and our method. As Table 3 shows, our method improves over OnlineAnySeg on all metrics with this backbone as well, which supports that probabilistic fusion is advantageous over hand-crafted hard fusion irrespective of the 2D models. A comparison with Table 1 also shows that class-agnostic AP is less affected by the choice of semantic descriptor.

Table 3: Backbone sensitivity: OnlineAnySeg and our method with SAM masks and CLIP descriptors in place of CropFormer and LSeg, with ground-truth poses.

Method	ScanNetV2					Replica				
	AP	mIoU	mAcc	PRQ (T)	PRQ (S)	AP	mIoU	mAcc	PRQ (T)	PRQ (S)
OnlineAnySeg	19.73	29.77	50.80	37.96	26.79	15.22	37.62	46.19	24.17	25.46
PartIMap(ours)	**21.93**	**32.91**	**54.33**	**39.54**	**28.30**	**16.00**	**38.84**	**47.70**	**26.45**	**26.19**

Component analysis. Table 4 analyzes the contribution of each proposed component on ScanNetV2, from OnlineAnySeg as the baseline (#0). Replacing baseline's descriptor averaging with vMF semantics while merging with vanilla probabilistic data association (#1, Appendix C) yields marginal changes: higher-order relations between masks are ignored, which cancels the gains of the vMF representation. (#2) Adding the supporter rule of OnlineAnySeg recovers these relations heuristically and improves all metrics over the baseline. Replacing both with the partition model (#3) improves all metrics further, which indicates that optimizing an explicit objective finds better groupings. Finally, the per-voxel language grid and the Beta occupancy (#4) allow the model to exploit fine-grained voxel-level information and give the best results.

Table 4: Ablation of the proposed components on ScanNetV2 with ground-truth poses. PDA: probabilistic data association; Supp.: baseline's supporter rule; Part.: partition model; Prob. rep.: vMF with per-voxel language grid and Beta occupancy.

Model		AP	mIoU	mAcc	PRQ (T)	PRQ (S)
#0	OnlineAnySeg (baseline)	22.5	45.33	65.85	47.90	42.36
#1	vMF + PDA	22.7	46.00	66.14	47.60	42.81
#2	vMF + PDA + Supp.	23.7	46.92	67.59	48.78	43.13
#3	vMF + Part.	<u>24.8</u>	<u>47.84</u>	<u>69.11</u>	<u>49.60</u>	<u>43.81</u>
#4	Prob. rep. + Part. (Ours)	**25.7**	**49.06**	**70.33**	**51.84**	**44.90**

Table 5: Sensitivity to the prior weight θ on ScanNetV2 with ground-truth poses. $\theta = 3$ is selected on a disjoint subset of the training split.

θ	AP	mIoU	mAcc	PRQ (T)	PRQ (S)
1	23.9	48.26	69.20	48.69	42.61
2	<u>25.2</u>	48.92	69.59	<u>49.90</u>	<u>44.73</u>
3	**25.7**	<u>49.06</u>	**70.33**	**51.84**	**44.90**
4	25.1	**49.54**	<u>69.79</u>	<u>49.90</u>	43.34
5	23.8	48.60	68.93	48.62	43.20

Hyperparameter. We fine-tune the partition model's prior weight θ on a 10-scene subset of the ScanNetV2 training split that is disjoint from our test set. Table 5 sweeps θ on the test scenes: the value selected on the training subset, $\theta = 3$, is also optimal on the evaluation scenes. The results vary smoothly around it, with AP staying within 1.9 points across the whole range. The system is therefore not sensitive to this hyperparameter and does not require exhaustive tuning.

4.4 Qualitative Results

Semantic segmentation. Figure 3 compares 3D semantic segmentation on four typical scenes with ground-truth poses. The online baselines OnlineAnySeg and OVI-MAP assign a single descriptor to each instance, so an incorrect or fragmented merge produces the scattered mislabeled patches visible in their results. Our method instead keeps semantic evidence at the voxel level and regularizes it with the instance consensus (Section 3.2), which yields coherent stuff regions, and preserves the labels of small and rare objects such as sinks and toilets. Our representation retains semantic and spatial ambiguity and the same behavior carries over to panoptic segmentation ; qualitative panoptic results and results under live SLAM are provided in Appendix F.

Visualizing uncertainty. The probabilistic representation also provides per-voxel uncertainty estimates. From the vMF statistics of the language grid (Sec. 3.2), we measure semantic uncertainty as $1 - \bar{r}_v$, which equals 0 when all observations agree and approaches 1 when they disagree. From the Beta occupancy (Eq. (2)), we measure spatial uncertainty as $1 - \mathbb{E}[\mathrm{occ}_v^k]$; a high value indicates that the voxel is supported by few or weakly weighted associations. Figure 4 visualizes both quantities next to the corresponding panoptic and semantic error maps. Semantic uncertainty is concentrated at the boundaries between instances and in the regions where the predicted class is wrong, so it correlates with semantic error. Occupancy uncertainty concentrates on the voxels of instances that are erroneously merged into or split from neighboring objects, and thus explains the remaining panoptic errors. Uncertainty visualizations under live SLAM are provided in Appendix F.

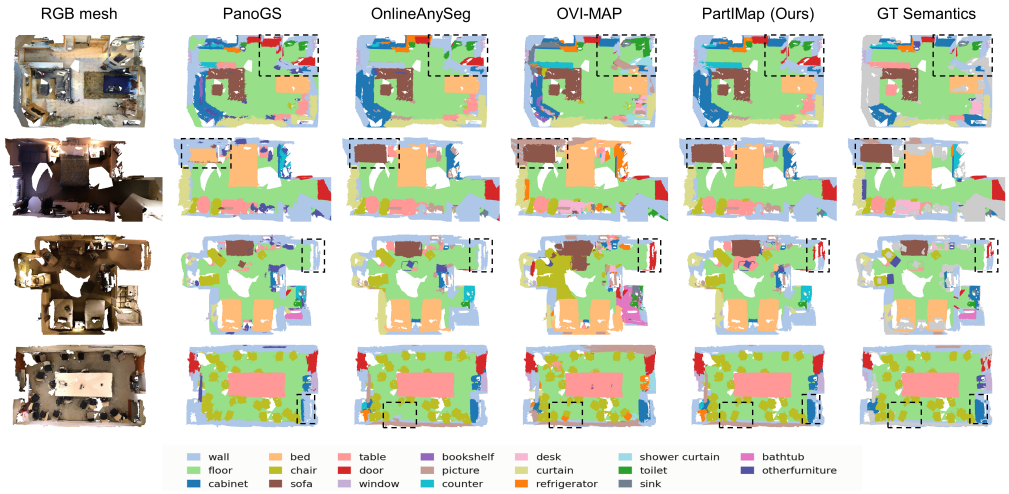

Figure 3: Qualitative 3D semantic segmentation on 4 typical scenes with ground-truth poses. From left to right: the reconstructed RGB mesh, the predictions of the offline method PanoGS and the online methods OnlineAnySeg and OVI-MAP, our prediction, and the ground truth.

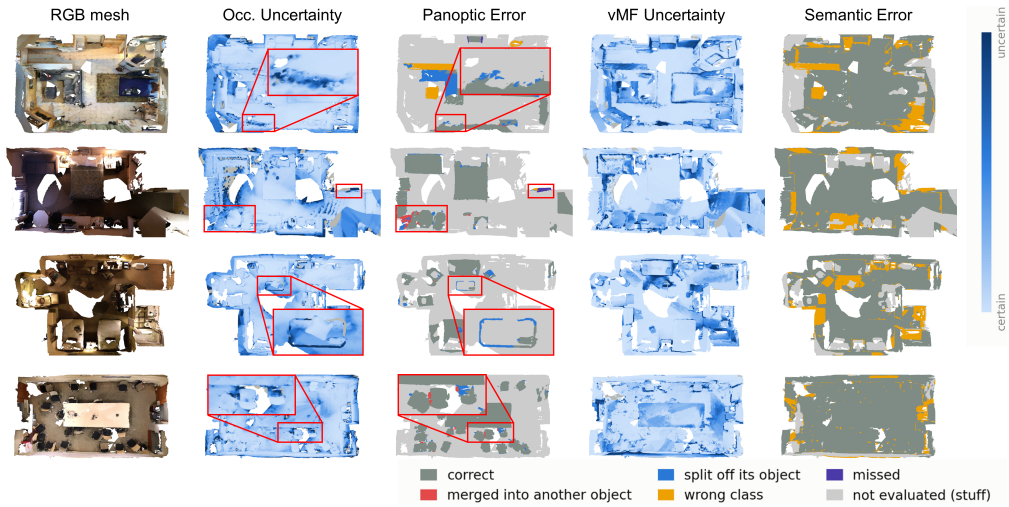

Figure 4: Uncertainty visualization on the same 4 scenes with ground-truth poses. From left to right: the reconstructed RGB mesh, spatial uncertainty, the panoptic error map, semantic uncertainty, and the semantic error map. Darker blue denotes higher uncertainty. Insets highlight regions where high occupancy uncertainty correlates with merge and split errors.

5 Conclusion

We presented PartIMap, an online open-vocabulary 3D panoptic mapping system that replaces the hand-designed hard rules of existing pipelines with inference under a single objective of a probabilistic model. A probabilistic voxel representation keeps semantics as von Mises–Fisher distributions and instance geometry as Beta occupancy, both defined by additive statistics. Mask merging is cast as maximum a posteriori inference of a partition on a mask graph, which retains probabilistic data association as its first-order relation while capturing the higher-order relations that previous systems approximated heuristically, and which yields soft membership weights from the same posterior. A per-observation provenance invariant makes the map re-derivable, so that PartIMap runs on a live SLAM system and re-fuses the map at loop closure. Without any training, PartIMap outperforms online and offline methods on ScanNetV2 and Replica, degrades the least under live SLAM poses, and provides qualitative per-voxel uncertainty.

AI USE STATEMENT

Large language models (LLMs) were used to assist with the preparation of this manuscript. Specifically, an LLM was used for language polishing, including grammar checking, sentence rephrasing, improving readability and clarity, and refining the overall flow of the text. The LLM was also used to assist in drafting initial versions of some sections of the manuscript. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of the paper. The LLM was not used to develop the research methodology, conduct experiments, analyze results, or formulate the paper's technical claims and conclusions.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. It involves no human subjects, user studies, crowdsourcing, or animal experimentation, and no new data were collected for this study.

Datasets and privacy. All experiments are conducted on two publicly available indoor RGB-D benchmarks, ScanNetV2 (Dai et al., 2017) and Replica (Straub et al., 2019), which were obtained through their official distribution channels and used in accordance with their respective terms of use for non-commercial research. ScanNetV2 contains scans of real indoor environments. We use only the released RGB-D sequences, camera poses, and annotations. We make no attempt to identify any individual or location from them, and we do not redistribute any dataset content. No personally identifiable information was used in this work.

Third-party models and code. PartIMap is zero-shot and requires no training. It relies on publicly released pretrained 2D foundation models, namely LSeg (Li et al., 2022), CropFormer (Qi et al., 2023), SAM (Kirillov et al., 2023), and CLIP (Radford et al., 2021), and it queries an unmodified ORB-SLAM3 (Campos et al., 2021) running as a separate process. Baselines were reproduced from their official public implementations. Results marked in Table 1 were taken from OnlinePG (Zhai et al., 2026), since no public code is available for OnlinePG. All external assets are used solely for academic research and benchmarking, with proper citation and in compliance with their licenses. Any code we release will preserve the required attributions and will not redistribute third-party models or data beyond their original license terms.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. The main paper provides the necessary details of the proposed method, experimental setup, and evaluation protocol. Additional implementation and experimental details are provided in the appendix and supplementary materials. Where applicable, the data processing procedures are also provided as supplementary materials to facilitate replication and verification of our results. Our codes will be released upon publication.

REFERENCES

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(46): 1345–1382, 2005.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine learning*, 56(1): 89–113, 2004.

Yaakov Bar-Shalom, Fred Daum, and Jim Huang. The probabilistic data association filter. *IEEE Control Systems Magazine*, 29(6):82–100, 2009.

Sean L. Bowman, Nikolay Atanasov, Kostas Daniilidis, and George J. Pappas. Probabilistic data association for semantic slam. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1722–1729, 2017.

Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8957–8965, 2019.

Manuel Dahnert, Ji Hou, Matthias Niessner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8282–8293, 2021.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, 2017.

Zilong Deng, Federico Tombari, Marc Pollefeys, Johanna Wald, and Daniel Barath. Ovi-map: Open-vocabulary instance-semantic mapping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12606–12616, 2026.

Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, Boris Ivanovic, Marco Pavone, and Yue Wang. Large spatial model: End-to-end unposed images to semantic 3d. In *Advances in Neural Information Processing Systems*, volume 37, pp. 40212–40229, 2024.

Margarita Grinvald, Fadri Furrer, Tonci Novkovic, Jen Jen Chung, Cesar Cadena, Roland Siegwart, and Juan Nieto. Volumetric instance-aware semantic mapping and 3d object discovery. *IEEE Robotics and Automation Letters*, 4(3):3037–3044, 2019.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028, 2024.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Ganesh Iyer, Soroush Saryazdi, Tao Chen, Alaa Maalouf, Shuang Li, Nikhil Varma Keetha, Ayush Tewari, Joshua Tenenbaum, Celso de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. ConceptFusion: Open-set multimodal 3D mapping. In *Proceedings of Robotics: Science and Systems*, July 2023.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19672–19682, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, 2023.

Seungjun Lee, Zihan Wang, Yunsong Wang, and Gim Hee Lee. Embodiedsplat: Online feed-forward semantic 3dgs for open-vocabulary 3d scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23774–23784, 2026.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

Haijie Li, Yanmin Wu, Jiarui Meng, Qiankun Gao, Zhiyao Zhang, Ronggang Wang, and Jian Zhang. Instancegaussian: Appearance-semantic joint gaussian representation for 3d instance-level perception. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14078–14088, 2025.

Hao Li, Zhengyu Zou, Fangfu Liu, Xuanyang Zhang, Fangzhou Hong, Yukang Cao, Yushi LAN, Manyuan Zhang, Gang YU, Dingwen Zhang, and Ziwei Liu. IGGT: Instance-grounded geometry transformer for semantic 3d reconstruction. In *International Conference on Learning Representations*, 2026.

Leyao Liu, Tian Zheng, Yun-Jou Lin, Kai Ni, and Lu Fang. Ins-conv: Incremental sparse convolution for online 3d segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18975–18984, 2022.

Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial Intelligence*, 293:103448, 2021.

R.P.S. Mahler. Multitarget bayes filtering via first-order multitarget moments. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1152–1178, 2003.

Tomas Berriel Martins, Martin R. Oswald, and Javier Civera. Open-vocabulary online semantic mapping for slam. *IEEE Robotics and Automation Letters*, 10(11):11745–11752, 2025.

Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian splatting slam. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18039–18048, 2024.

John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4628–4635, 2017.

John Mccormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: Volumetric object-level slam. In *International Conference on 3D Vision (3DV)*, pp. 32–41, 2018.

Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2901–2910, 2019.

Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4205–4212, 2019.

Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4018–4028, 2024.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824, 2023.

Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4047–4056, 2023.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20051–20060, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

D. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24 (6):843–854, 1979.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Michael Strecke and Jorg Stuckler. Em-fusion: Dynamic object-level slam with probabilistic data association. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Xiangyu Sun, Haoyi Jiang, Liu Liu, Seungtae Nam, Gyeongjin Kang, Xinjie Wang, Wei Sui, Zhizhong Su, Wenyu Liu, Xinggang Wang, and Eunbyung Park. Uni3r: Unified 3d reconstruction and semantic understanding via generalizable gaussian splatting from unposed multi-view images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33280–33290, 2026.

Ayca Takmaz, Elisabetta Fedele, Robert Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68367–68390, 2023.

Yijie Tang, Jiazhao Zhang, Yuqing Lan, Yulan Guo, Dezun Dong, Chenyang Zhu, and Kai Xu. Onlineanyseg: Online zero-shot 3d segmentation by visual foundation model guided 2d mask merging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3676–3685, 2025.

Hanshi Wang, Cai Zijian, Jin Gao, Yiwei Zhang, Weiming Hu, Ke Wang, and Zhipeng Zhang. Online segment any 3d thing as instance tracking. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 49076–49095, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024.

Joey Wilson, Yuewei Fu, Joshua Friesen, Parker Ewen, Andrew Capodieci, Paramsothy Jayakumar, Kira Barton, and Maani Ghaffari. Convbki: Real-time probabilistic semantic mapping network with quantifiable uncertainty. *IEEE Transactions on Robotics*, 40:4648–4667, 2024.

Xiuwei Xu, Chong Xia, Ziwei Wang, Linqing Zhao, Yueqi Duan, Jie Zhou, and Jiwen Lu. Memory-based adapters for online 3d scene perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21604–21613, 2024.

Xiuwei Xu, Huangxing Chen, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Embodiedsam: Online segment any 3d thing in real time. In *International Conference on Learning Representations*, pp. 38020–38035, 2025.

Kashu Yamazaki, Taisei Hanyu, Khoa Vo, Thang Pham, Minh Tran, Gianfranco Doretto, Anh Nguyen, and Ngan Le. Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9411–9417, 2024.

Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28274–28284, 2024.

Yi Yang, Myrna Castillo, Bodo Rosenhahn, and Michael Ying Yang. Puf: Plug-and-play uncertainty-aware fusion for online 3d scene graph generation. In *European Conference on Computer Vision (ECCV)*, pp. 462–479, 2026.

Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision (ECCV)*, pp. 162–179, 2024.

Hongjia Zhai, Hai Li, Zhenzhe Li, Xiaokun Pan, Yijia He, and Guofeng Zhang. Panogs: Gaussian-based panoptic segmentation for 3d open vocabulary scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14114–14124, 2025.

Hongjia Zhai, Qi Zhang, Xiaokun Pan, Xiyu Zhang, Yitong Dong, Huaqi Zhang, Dan Xu, and Guofeng Zhang. Onlinepg: Online open-vocabulary panoptic mapping with 3d gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33269–33279, 2026.

Runsong Zhu, Jiaxin Guo, Xiaoyang Guo, Zhengzhe Liu, Ka-Hei Hui, Wei Yin, Kai Chen, Wei Chen, Weiqiang Ren, Yunhui Liu, Pheng-Ann Heng, and Chi-Wing Fu. EPS3d: End-to-end feed-forward 3d panoptic segmentation. In *International Conference on Machine Learning*, 2026.

APPENDIX

The appendix is organized as follows.

- Appendix A defines the mask bank and the mask graph on which the partition model operates, and gives the online pipeline and the loop-closure re-fusion as pseudocode (Algorithms 1 and 2).
- Appendix B derives the posterior, the merge gain, and the membership weights of the partition model from its likelihood and Gibbs prior.
- Appendix C relates the partition model to probabilistic data association and shows which higher-order relations a per-node matching cannot represent.
- Appendix D lists all constants and details the semantic cue calibration, the anchoring of ORB-SLAM3 poses, and the handling of map resets and merges.
- Appendix E isolates the effect of the loop-closure re-fusion under live SLAM.
- Appendix F presents additional qualitative results for panoptic segmentation, semantic segmentation under live SLAM, and uncertainty under live SLAM.
- Appendix G discusses limitations and future work.

A MASK FUSION: REPRESENTATION AND ONLINE PIPELINE

This section describes the data structures that hold observations and instances, and the complete online procedure that the main text summarizes. Section A.1 defines the mask bank and the mask graph on which the partition model of Section 3.3 operates, following the organization of (Tang et al., 2025; Zhai et al., 2026). Section A.2 gives the full pipeline under ground-truth and live SLAM poses as pseudocode.

A.1 MASK BANK AND MASK GRAPH

Observations. Every keyframe is segmented by the class-agnostic 2D segmenter, and each mask is back-projected through the depth map into the voxel grid. Each surviving mask becomes an observation o with: a global ID, its source keyframe, its voxel set V_o in world-frame voxel coordinates, its unit descriptor $\boldsymbol{x}_o$, and its provenance weight β_o, initialized to 1. Observations are appended to a bank and never deleted; an observation whose instance is removed keeps its record with $\beta_o = 0$.

Hashed voxel volume and mask index. The scene is reconstructed as a TSDF in a hashed voxel-block grid, with 4^3 voxels per block. In addition to the TSDF value, each voxel entry stores the set of IDs of the raw observations whose voxel set covers it. This index is append-only: it is written when an observation is inserted, and is not rewritten when instances merge. Alongside it, a per-frame record stores the set of voxels back-projected from the depth map of each keyframe, which defines visibility.

Instances and the mapping table. An instance k is a set of observations, together with its voxel set $V_k = \bigcup_{o \in k} V_o$ and the fused statistics of Eqs. (1) and (2). A mapping table translates every observation ID to the ID of the instance that currently holds it, or to a null value once removed, and each instance keeps the inverse list of its constituent IDs. Merging instances updates only this mapping and the fused statistics; every containment query translates the raw IDs found in the voxel index through the mapping at read time, filtering removed observations. Instances are therefore a view over the raw-observation layer, which makes merging efficient.

Visibility and containment. A voxel is visible in a keyframe if it belongs to that keyframe's back-projected voxel set. The system maintains a voxel $\times$ keyframe visibility matrix and an instance $\times$ keyframe matrix recording which keyframes observed each instance. Let Vis_a denote the union of the voxels visible in the keyframes that observed instance a. Instance b counts as visible to a if at least 20% of its voxels, or at least 500 voxels, lie in Vis_a; otherwise c_{ab} is undefined and set to zero.

For visible pairs, the visibility-gated containment ratio is

$$c_{ab} = \frac{|V_a \cap V_b|}{|V_b \cap \mathrm{Vis}_a|}, \tag{9}$$

the fraction of b that a contains, counted only over the part of b that the keyframes of a could have seen; values below 0.025 are set to zero. The containment matrix over all instances is maintained incrementally: the rows and columns of newly inserted or merged instances are recomputed by querying the voxel index, while the others are kept. Under loop closure it is recomputed from scratch (Algorithm 2).

Mask graph. At every merge step, the mask graph has one node per current instance, including the observations inserted since the previous step as singleton instances. An edge (a, b) carries the evidence λ_{ab} of Eq. (6) if $c_{ab} > 0$, and no evidence otherwise. Pairs that were never observed together carry no information (i.e., $\lambda_{ab} = 0$). Because containment is nonzero only for instances that share voxels within a co-visible region, the graph decomposes into small connected components, each of which is an independent instance of the partition problem of Section 3.3.

A.2 Online pipeline

Algorithm 1 lists the per-frame loop. With ground-truth poses, it constitutes the whole system. With live SLAM, poses are queried from the tracker for every frame, keyframe corrections are polled after every frame, and a reported loop closure triggers Algorithm 2 before the frame is processed; the re-merge inside Algorithm 2 is the same mask graph partition step as in the normal loop. Depth integration and mask insertion take place on every 10th frame, and a partition step is run every 5 inserted keyframes. All remaining constants are listed in Appendix D.

B Derivation of the partition model

Prior. The model of Section 3.3 places a Gibbs prior on the partition X of the nodes of the mask graph,

$$p(X) = \frac{1}{Z_\theta} \exp\Big(-\theta \sum_{(a,b)\in\mathcal{E}} X_{ab}\Big), \qquad Z_\theta = \sum_{X'} \exp\Big(-\theta \sum_{(a,b)\in\mathcal{E}} X'_{ab}\Big), \tag{10}$$

where the sum in Z_θ runs over all partitions of the n nodes. The set of partitions is finite, so the prior is proper for every θ. For a single edge in isolation, $p(X_{ab} = 1)/p(X_{ab} = 0) = e^{-\theta}$, so θ is the negative prior log-odds that two nodes connected by an edge are the same object. Pairs outside $\mathcal{E}$ do not enter the exponent.

Posterior. The pairwise observations y_{ab} on the edges are conditionally independent given X, Eq. (3). Each factor can be written relative to the $X_{ab} = 0$ baseline,

$$p(y_{ab} \mid X_{ab}) = p(y_{ab} \mid X_{ab} = 0)\; e^{\lambda_{ab} X_{ab}}, \tag{11}$$

which holds for both values of X_{ab} by the definition of λ_{ab}. Taking the product over all edges, the baseline $\prod_{(a,b)\in\mathcal{E}} p(y_{ab} \mid X_{ab} = 0)$ does not depend on X, and neither does Z_θ, so

$$\log p(X \mid Y) = \log p(Y \mid X) + \log p(X) + \text{const} = \sum_{(a,b)\in\mathcal{E}} \lambda_{ab} X_{ab} - \theta \sum_{(a,b)\in\mathcal{E}} X_{ab} + \text{const}, \tag{12}$$

which is Eq. (4).

Merge gain. Let X^{sep} be a partition in which P and Q are distinct clusters, and X^{mrg} the partition obtained from it by merging P and Q. The two differ only on the entries (a, b) with $a \in P$ and $b \in Q$, which are 0 in X^{sep} and 1 in X^{mrg}, and of these only the edges $\mathcal{E}_{PQ}$ appear in Eq. (4). Hence

$$\log p(X^{\mathrm{mrg}} \mid Y) - \log p(X^{\mathrm{sep}} \mid Y) = \sum_{(a,b)\in\mathcal{E}_{PQ}} \big(\lambda_{ab} - \theta\big), \tag{13}$$

which is Eq. (7). The gain is positive if and only if the mean evidence $\bar{\lambda}_{PQ}$ over $\mathcal{E}_{PQ}$ exceeds θ. The accept/reject test of the greedy estimator is therefore that of average-linkage agglomerative clustering with threshold θ, with the average taken over edges only, and Eq. (4) is the objective of which this test is the greedy step.

Algorithm 1 Online mapping loop of PartIMap (ground-truth or live SLAM poses)

Require: RGB-D stream $\{(I_t, D_t)\}$; poses $\boldsymbol{T}_t$, given or queried from the tracker; keyframe interval Δ_{kf}; merge interval Δ_{m}

initialize the TSDF volume, the observation bank $\mathcal{B}$, the instance set $\mathcal{I}$, and the language grid
for each frame t **do**
  if live SLAM **then**
    query $\boldsymbol{T}_t$ and poll keyframe corrections; **if** a loop closure is reported **then** run Algorithm 2
  end if
  if $t \bmod \Delta_{\mathrm{kf}} \neq 0$ **then continue**
  end if
  integrate D_t into the TSDF under $\boldsymbol{T}_t$; record the back-projected voxel set of frame t (visibility)
  $\{m_i\} \leftarrow$ class-agnostic 2D masks of I_t; $\{\boldsymbol{x}_i\} \leftarrow$ masked mean-pooled descriptors of the vision-language model
  for each mask m_i passing the size filters **do**
    $V_o \leftarrow$ back-projection of m_i through D_t; append $o = (V_o, \boldsymbol{x}_o, t)$ with $\beta_o = 1$ to $\mathcal{B}$
    insert the id of o into the voxel index at V_o; scatter $\boldsymbol{x}_o$ into the language grid, Section 3.2
    create a singleton instance with $\boldsymbol{R} = \boldsymbol{x}_o$, $W = 1$, $a_v = \mathbf{1}[v \in V_o]$; map o to it
  end for
  update the containment rows and columns of the new instances, Eq. (9)
  if $t \bmod (\Delta_{\mathrm{kf}}\Delta_{\mathrm{m}}) = 0$ **then**
    GRAPHPARTITIONSTEP($\mathcal{I}$)
  end if
end for

procedure GRAPHPARTITIONSTEP($\mathcal{I}$)
  compute s_{ab} and λ_{ab} for all pairs with $c_{ab} > 0$, Eqs. (5) and (6); all other $\lambda_{ab} = 0$
  for each connected component of the evidence graph **do**
    start from singletons; repeatedly merge the pair (P, Q) with the largest $\Delta(P, Q)$ of Eq. (7) while $\Delta > 0$
  end for
  for each resulting cluster $\mathcal{C}$ with $|\mathcal{C}| \geq 2$ **do**
    compute β_a for every member $a \in \mathcal{C}$, Eq. (8), with $\beta = 1$ for the seed
    multiply the provenance weight β_o of every observation of a by β_a
    fuse statistics: $\boldsymbol{R} \leftarrow \sum_a \beta_a \boldsymbol{R}_a$, $W \leftarrow \sum_a \beta_a W_a$, $a_v \leftarrow \sum_a \beta_a a_v^a$; $V \leftarrow \bigcup_a V_a$
    map the observations of all members to the merged instance
  end for
  update the containment rows and columns of the merged instances
end procedure

Algorithm 2 Loop-closure re-fusion

Require: corrected keyframe poses $\{\boldsymbol{T}'_{\mathrm{kf}}\}$ from the tracker; observation bank $\mathcal{B}$; instance set $\mathcal{I}$

1. **Propagate:** for every integrated frame, compose its relative pose to the nearest corrected keyframe with that keyframe's correction, taken incrementally with respect to the last applied correction
2. **Re-integrate:** build a new TSDF volume by replaying every integrated depth map under the corrected poses; swap to the new TSDF
3. **Re-warp:** for every observation $o \in \mathcal{B}$, transform V_o by $\boldsymbol{T}'_{\mathrm{kf}}\boldsymbol{T}_{\mathrm{kf}}^{-1}$ of its keyframe, snap to the lattice, and drop voxels whose block the new volume did not allocate; transform the per-frame visibility sets similarly
4. **Rebuild:** clear and re-insert the voxel index from the re-warped V_o; rebuild the visibility matrices; for every instance k set $V_k \leftarrow \bigcup_{o \in k} V_o$ and re-derive $(\boldsymbol{R}_k, W_k, a_v^k)$ from Eqs. (1,2); re-scatter the language grid from $\{(\boldsymbol{x}_o, V_o)\}$ ▷ mapping table and β_o are untouched
5. **Recompute** the full containment matrix from the corrected geometry, Eq. (9)
6. GRAPHPARTITIONSTEP($\mathcal{I}$) ▷ one ordinary merge step, Algorithm 1

Membership weight. Let $\hat{X}$ be the partition returned by the estimator and $\mathcal{C} = \mathcal{C}(a)$ the cluster containing node a, with $|\mathcal{C}| \geq 2$. We contrast two hypotheses: $X^{\text{in}} := \hat{X}$, in which a is a member of $\mathcal{C}$, and X^{out}, which equals $\hat{X}$ except that a is moved into its own singleton, i.e., $X^{\text{out}}_{ab} = 0$ for all $b \in \mathcal{C} \setminus a$. All other entries are identical in both partitions. We define the membership weight as the probability of X^{in} under these two hypotheses,

$$\beta_a := \frac{p(X^{\text{in}} \mid Y)}{p(X^{\text{in}} \mid Y) + p(X^{\text{out}} \mid Y)} = \sigma\big(\log p(X^{\text{in}} \mid Y) - \log p(X^{\text{out}} \mid Y)\big), \tag{14}$$

where the second equality uses $e^u/(e^u + e^v) = \sigma(u - v)$. Evaluating the difference with Eq. (4), the two partitions differ only on the entries (a, b) with $b \in \mathcal{C} \setminus a$, which are 1 in X^{in} and 0 in X^{out}, and of these only the edges with $b \in \mathcal{N}(a)$ enter the posterior:

$$\log p(X^{\text{in}} \mid Y) - \log p(X^{\text{out}} \mid Y) = \sum_{b \in \mathcal{C} \cap \mathcal{N}(a)} \big(\lambda_{ab} - \theta\big). \tag{15}$$

Substituting into Eq. (14) gives Eq. (8). Because the prior depends only on the asserted edges, moving a out of $\mathcal{C}$ does not change other terms, and the membership weight is the merge gain of Eq. (7) evaluated for the pair $(\{a\}, \mathcal{C} \setminus a)$.

Notably, the partition determines which nodes form an object, whereas β_a determines how strongly each member contributes, so that a marginal merge affects the descriptor and occupancy of the object only in proportion to its evidence.

C Relation to probabilistic data association

Single-node conditional. Consider a node a and fix the identities z_b of all other nodes, which partition them into clusters $\{\mathcal{C}_j\}_{j=1}^{M}$. Under Eq. (4), assigning a to cluster j changes the log-posterior by $\Lambda_{aj} := \sum_{b \in \mathcal{C}_j \cap \mathcal{N}(a)} (\lambda_{ab} - \theta)$, whereas placing a in a new singleton cluster leaves it unchanged. The conditional posterior of the identity of a is therefore the softmax

$$p(z_a = j \mid Y,\, z_{\setminus a}) = \frac{e^{\Lambda_{aj}}}{Z_a}, \qquad p(z_a = \text{new} \mid Y,\, z_{\setminus a}) = \frac{1}{Z_a}, \qquad Z_a = 1 + \sum_{j'=1}^{M} e^{\Lambda_{aj'}}. \tag{16}$$

A cluster without an edge to a has $\Lambda_{aj} = 0$. Restricting the sum to the clusters adjacent to a plays the role of the validation gate of data association. When each adjacent instance is represented by a single node j with $\lambda_{aj} = T(s_{aj} - s_{\text{stay}})$ from Eq. (6), Eq. (16) reduces to

$$\beta_{aj} = \frac{e^{T s_{aj}}}{e^{T s_{\text{stay}} + \theta} + \sum_{j'} e^{T s_{aj'}}}, \tag{17}$$

which is the tempered-softmax association weight of probabilistic data association with an explicit stay/birth alternative at indifference level $s_{\text{stay}} + \theta/T$, in the spirit of the joint probabilistic data association filter (Bar-Shalom et al., 2009). The partition model therefore contains probabilistic association as its first-order relation, with the same score and slope, and a greedy per-node assignment that thresholds Eq. (17) at $\frac{1}{2}$ recovers the data-association merge rule.

Matching versus partition. The weights of Eq. (17) are the marginals of a matching, $\sum_j \beta_{aj} \leq 1$, and answer the question of which single instance a belongs to. The latent variable of the merging problem, however, is a partition. In the case of near-duplicated masks generated by oversegmentation, the partition returns all of them. Probabilistic data association's row normalization divides one unit of mass among all the masks. Consider $k \geq 3$ mutual duplicates with equal pairwise evidence $\lambda > \theta$: for each of them, the weight of any other member is

$$\beta = \frac{e^{\lambda - \theta}}{1 + (k-1)\, e^{\lambda - \theta}} < \frac{1}{k-1} \leq \frac{1}{2}, \tag{18}$$

so no pair exceeds the threshold regardless of how large λ is, and the mask group is never merged. Under Eq. (4), by contrast, adding one more duplicate to a cluster of i duplicates has gain $i(\lambda - \theta)$ by Eq. (7), which is positive whenever $\lambda > \theta$. This is the condition a two-node merge must satisfy

as well, so the greedy estimator collects the whole group into one cluster, on a criterion that does not depend on k. The same joint treatment yields transitivity: for a chain a–c–b with $\lambda_{ac}, \lambda_{bc} > \theta$ and no edge between a and b, the partition $\{a, b, c\}$ has log-posterior gain $(\lambda_{ac} - \theta) + (\lambda_{bc} - \theta)$ over singletons and dominates both $\{a, c\}, \{b\}$ and $\{a\}, \{b, c\}$, so the three nodes are merged without direct evidence between a and b. A per-node matching cannot represent such higher-order relations, which is why prior online systems (Yan et al., 2024; Tang et al., 2025) add a heuristic supporter rule on top of the pairwise test.

D IMPLEMENTATION DETAILS

Hyperparameters and constants. Table 6 lists all constants of the system. The link parameters T and s_{stay} were selected for the data-association stage on the ScanNetV2 training subset and inherited unchanged by the partition model, whose only tuned parameter is the prior weight θ (Section 4.3). The voxel size, mask size and visibility thresholds are inherited from (Tang et al., 2025; Zhai et al., 2026).

Table 6: Constants of PartIMap. Values are identical on ScanNetV2 and Replica unless stated.

Group	Quantity	Value
Cadence	TSDF integration and mask insertion	every 10th frame
	partition step	every 5 inserted keyframes
	unsupported-instance removal	≤ 1 observation and no merge for 20 keyframes
Voxel grid	voxel size / block size	2.5 cm / 4^3 voxels
	block capacity	100k (160k with live SLAM)
	TSDF truncation	10 cm
	depth range	0.1–5 m (ScanNetV2), 0.1–8 m (Replica)
Mask filtering	minimum 2D mask size / back-projected size	400 px / 20 voxels
	visibility gate	$\geq 20\%$ of voxels or ≥ 500 voxels
	containment floor	0.025
Association	slope T / indifference level s_{stay}	4.0 / 2.0
	prior weight θ	3.0
Representation	occupancy pseudo-count b_0	1.0
	decoding blend (Section 3.2)	instance resultant added with unit weight

Semantic cue calibration. The raw cosine similarity between vMF means is passed through a fixed monotone rescaling that maps the $[p_1, p_{99}]$ range of pairwise cosines in the descriptor space to $[0, 1]$. This unit conversion allows the same association parameters T and s_{stay} to serve different 2D back-ends whose descriptor spaces occupy different cosine ranges. Our hyperparameters therefore transfers to the backbone sensitivity study (Table 3) without re-tuning.

Tracking and anchoring. Camera poses are provided by an unmodified ORB-SLAM3 (Campos et al., 2021) in RGB-D mode, run as a separate process and queried once per frame. The ORB-SLAM3 world frame is aligned to the map frame once, at the first successfully tracked frame. No whole-trajectory alignment is performed, so the residual anchoring error is included in the reported live results. Keyframe corrections are polled after every frame. Each correction is converted into per-integrated-frame corrections as an incremental delta relative to the *last applied* corrected pose of that keyframe, so that successive loop closures compose rather than compound.

Guards. After an ORB-SLAM3 map reset, the tracker starts a new map in its atlas whose world frame no longer coincides with that of our map, and integrating the frames tracked in it would write misregistered geometry. The existing method that supports live SLAM, OVO (Martins et al., 2025), typically abandons the rest of the sequence. We implemented a fix for map reset. Integration and loop-closure re-fusion are suspended at a reset, but instead of abandon, frames keep feeding ORB-SLAM3, and once the camera revisits previously mapped space, ORB-SLAM3 merges the new map into the earlier one, and its poses return to the world frame of that map. To follow resets and merges, every integrated frame records the map in which it was tracked, each map keeps its own alignment to the map frame, and a keyframe correction is applied to the integrated frames of the same map. On a merge, the alignment reverts to that of the surviving map, and the frames recorded under the absorbed map are reassigned to it, so that their recorded poses serve as the reference for

the incremental corrections described above. Our fix successfully recovers multi-map ScanNetV2 scenes 200, 400, 590 and 645, which was not usable by original OVO. For a fair comparison, in Table 2 we applied the same fix to PartIMap, OVO and OnlineAnySeg.

E Additional ablation study

Loop-closure re-fusion. We isolate the effect of the loop-closure re-fusion of Section 3.4 by running our method on ScanNetV2 under four pose settings. *Live* uses the live ORB-SLAM3 poses but leaves the map unchanged at loop closure, so that the geometry integrated under drifted poses is retained. *Live+LC* is our full system as reported in Table 2, which re-fuses the map at every loop closure with Algorithm 2. *Replay* runs the method offline on the final loop-closure-corrected poses, as for the replayed baselines in Table 2. *GT* uses ground-truth poses. Replay and GT are not attainable online and serve as references: the gap between GT and Replay measures the residual error of the SLAM trajectory itself, whereas the gap between Replay and Live+LC measures the cost of deciding merges online before the trajectory is corrected.

Table 7: Ablation of the loop-closure re-fusion on ScanNetV2. *Live*: live ORB-SLAM3 poses without re-fusion at loop closure. *Live+LC*: live ORB-SLAM3 poses with loop-closure re-fusion (ours, as in Table 2). *Replay*: offline run on the final loop-closure-corrected ORB-SLAM3 trajectory. *GT*: ground-truth poses (as in Table 1). Replay and GT are not available online and are reported as references.

Input	AP	mIoU	mAcc	PRQ (T)	PRQ (S)
Live	17.2	45.51	65.48	45.41	37.33
Live+LC (ours)	20.0	48.26	67.23	49.78	39.67
Replay	23.2	48.84	68.59	49.96	41.25
GT	25.7	49.06	70.33	51.84	44.90

As Table 7 shows, loop-closure re-fusion improves all metrics over the Live setting. Relative to Replay, it recovers 96% of the gap in PRQ (T) and 83% of the gap in mIoU. The semantic metrics are recovered almost completely because the language grid is re-scattered from the stored per-observation descriptors under the corrected geometry, independently of the instance decisions. Replay, in contrast, never observes the drifted trajectory, which confirms that the Replay setting in Table 2 is favorable for the baselines evaluated in it. Finally, Replay remains below GT on all metrics, most notably by 3.65 PRQ (S), which reflects the residual pose and anchoring error of ORB-SLAM3 (Appendix D) that no map-side correction can remove.

F Additional qualitative results

Panoptic segmentation. Figure 5 shows qualitative 3D panoptic segmentation results on the same scenes as Figure 3 with ground-truth poses, complementing the semantic results. Baseline methods commonly over-segments large objects, for example OVI-MAP splits the conference table in the fourth scene into several patches. Such objects are covered by many partially overlapping masks, of which only some pairs are directly related by containment. Our partition model merges the multiple-view masks jointly through transitivity. PartIMap also does not over-merge adjacent objects of the same class. For example, the densely placed chairs around the table in the fourth scene remain separate instances. This agrees with our superior performance in PRQ (T) in Table 1.

Live SLAM. Figure 6 shows semantic segmentation results with poses estimated by ORB-SLAM3. Although PanoGS and OVI-MAP operate on the final corrected trajectory, pose error visibly degrades their predictions, especially on object boundaries. Our predictions remain close to those with ground-truth poses in Figure 3, in line with the small decline in mIoU from 49.06 to 48.26 (Tables 1 and 2). Because the language grid is re-scattered from the stored descriptors at every loop closure, the voxel-level semantics are re-derived under the corrected geometry (Appendix E).

Uncertainty under live SLAM. Figure 7 shows the corresponding uncertainty maps under live SLAM. Compared with ground-truth poses (Figure 4), occupancy uncertainty increases along walls

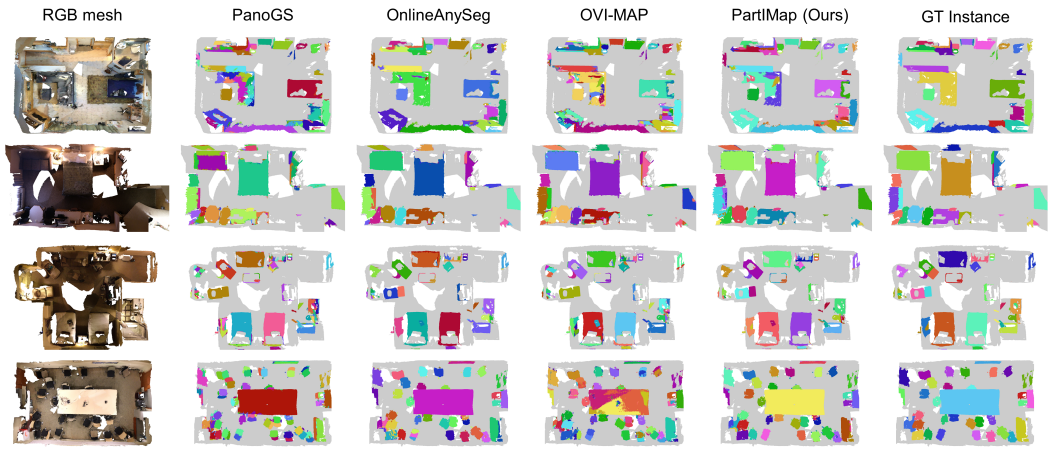

Figure 5: Qualitative 3D instance segmentation on the scenes of Figure 3 with ground-truth poses. From left to right: the reconstructed RGB mesh, the predictions of PanoGS, OnlineAnySeg, OVI-MAP, and our method, and the ground truth. Instances are shown in random colors; stuff regions are shown in gray.

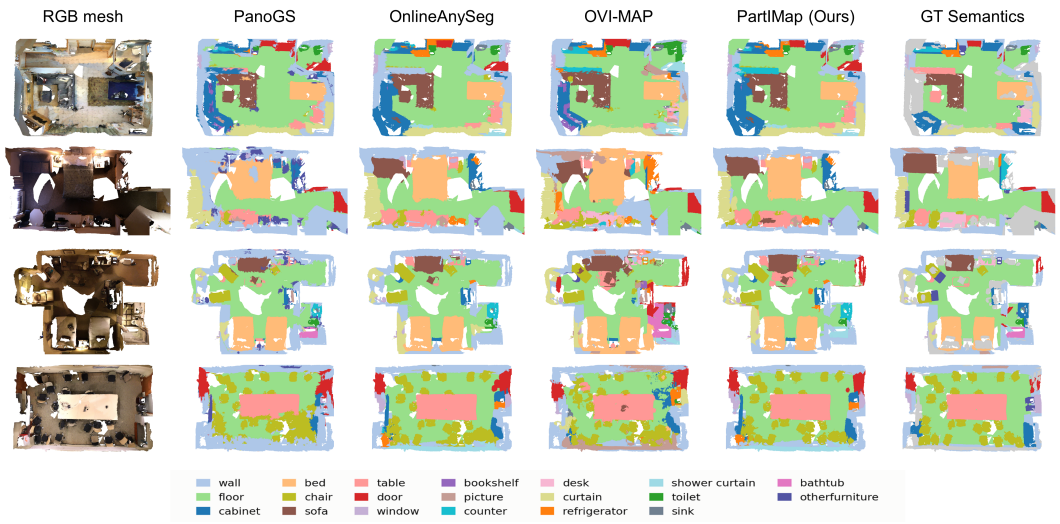

Figure 6: Qualitative 3D semantic segmentation on the scenes of Figure 3 with poses estimated by live SLAM, in the layout of Figure 3. Following Table 2, PanoGS and OVI-MAP are run offline on the final loop-closure-corrected trajectory (*Replay*), and OnlineAnySeg is equipped with our loop-closure re-fusion (*Live+LC*).

and object boundaries. The panoptic error maps show a corresponding increase in split errors. The high-uncertainty regions thus continue to coincide with these errors, which shows that the uncertainty estimates remain informative when the poses are noisy.

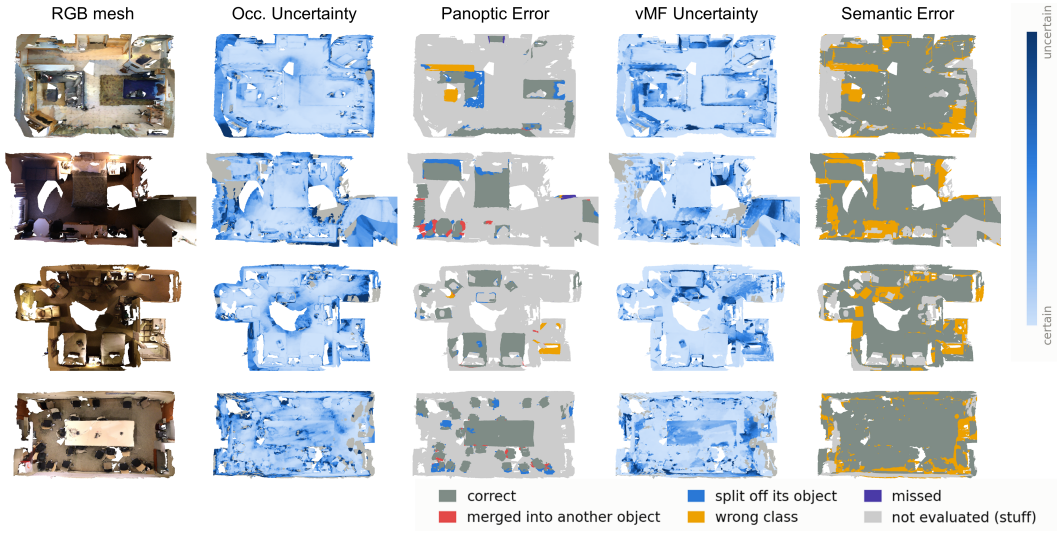

Figure 7: Uncertainty visualization on ScanNetV2 with poses estimated by live SLAM, in the layout of Figure 4.

G Limitations and Future Work

PartIMap assumes a static scene and requires depth input. Our uncertainty estimation remains qualitative, as the 2D backbones LSeg Li et al. (2022) and CLIP Radford et al. (2021) are deterministic models, and switching to their calibrated probabilistic variants goes beyond the scope of this research. Future work includes learning the pairwise evidence from data while keeping the partition model, extending the representation to dynamic objects, studying quantitative correlation between calibrated uncertainty and model error, and combining the probabilistic map with feed-forward geometry models to remove the dependence on depth sensing.