FROM VISUAL GEOMETRY EVIDENCE TO EMBODIED SEMANTIC OCCUPANCY MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents exploring indoor environments require reliable semantic occupancy memory that persists across observations and revisits. Building such memory is challenging because each observation provides incomplete and uncertain geometric and semantic evidence. We introduce GEM-Occ, a Gaussian Evidence Memory framework that consolidates evidence accumulated over time into persistent semantic occupancy memory. Local predictions are converted into occupied semantic Gaussians and free-space ray evidence. Confidence- and visibility-aware causal updates integrate supporting observations, suppress occupancy contradicted by observed free space, and preserve previously observed structures through occlusion. A hierarchical memory organization supports continued mapping and efficient queries across connected indoor spaces. To evaluate this capability, we introduce HIOcc, a unified benchmark for embodied semantic occupancy memory. HIOcc establishes a shared semantic label space and evaluation framework spanning local prediction, room-level online mapping, and building-level mapping, while accommodating perspective and panoramic observations. Experiments on HIOcc demonstrate that GEM-Occ outperforms existing methods, enabling accurate semantic occupancy prediction and consistent online mapping across spatial scales with efficient memory usage and fast occupancy queries.

1 INTRODUCTION

Embodied semantic occupancy mapping requires an agent to build a persistent spatial representation from observations acquired during exploration. Semantic occupancy provides a natural basis for this representation by describing occupied structures and their semantic categories (Song et al., 2017; Cao & De Charette, 2022; Wu et al., 2025b). As the agent moves between rooms, previously observed structures leave the current view but remain relevant to its understanding of the environment. Returning to these regions provides new evidence that may reinforce earlier estimates or reveal errors caused by incomplete views and uncertain predictions. A useful occupancy memory must therefore retain spatial knowledge beyond the current observation while remaining open to revision. This requires distinguishing occupied structures, observed free space, and unknown regions, and maintaining these distinctions as observations accumulate across connected indoor spaces.

Recent voxel-based and Gaussian-based methods have improved local semantic occupancy prediction (Cao & De Charette, 2022; Yu et al., 2024; Qian et al., 2026; Zhou et al., 2026a), while online methods extend perception through persistent scene representations. EmbodiedOcc (Wu et al., 2025b) maintains a global Gaussian memory and progressively refines regions within the current view. Its uniform initialization across a predefined scene extent, however, allocates memory to regions before they are observed, with the initial cost growing with scene volume. Incremental Gaussian fusion provides another route to constructing global maps from successive predictions (Zhou et al., 2026a). For continued exploration, memory construction must address both where new elements are allocated and how accumulated estimates are revised. Supporting observations should consolidate an existing estimate, observed free space should suppress contradicted occupancy, and occluded regions should retain their state in the absence of new evidence. These cases can occur at the same location at different times, making observation confidence and visibility central to memory updates. As the explored area grows, the accumulated representation must also support efficient access for subsequent updates and occupancy queries.

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.

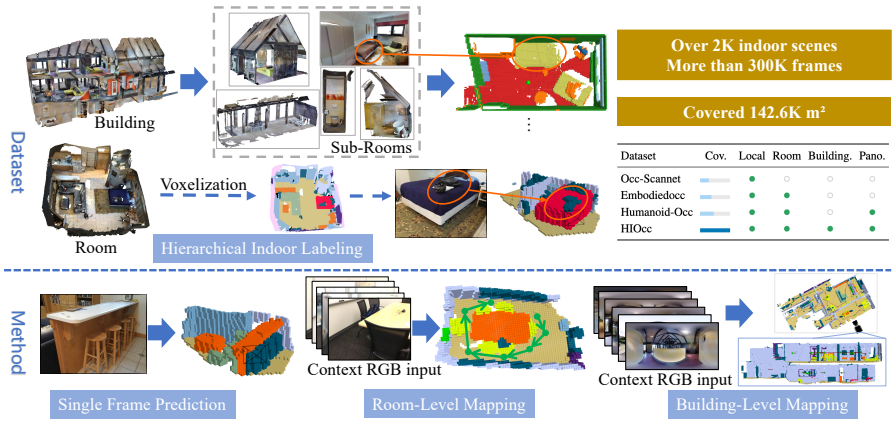

Figure 1: **Overview of HIOcc and GEM-Occ.** HIOcc provides hierarchical indoor semantic occupancy annotations across local views, rooms, and connected buildings. GEM-Occ builds persistent semantic Gaussian memory for local prediction, room-level online mapping, and building-level mapping.

We introduce GEM-Occ, a Gaussian Evidence Memory framework in which incoming observations determine both memory allocation and state updates. For each posed observation, predicted geometry and semantics are converted into occupied semantic Gaussians and free-space ray evidence in a shared world coordinate system. Geometry and uncertainty determine the spatial support of the Gaussians, while visible ray segments identify space observed to be empty. Incoming Gaussians are associated with existing memory elements or inserted to represent newly observed structure. Confidence-weighted fusion consolidates geometric and semantic estimates, and free-space evidence revises contradicted occupancy without treating occlusion as evidence of absence. Spatial submaps and a building-level connectivity graph organize the growing memory, while merging and pruning consolidate redundant or poorly supported elements. The resulting representation supports occupancy queries throughout exploration. Training supervises both local predictions and fused memory outputs, connecting observation-level evidence construction with the quality of the accumulated map.

Evaluating this capability requires measuring how map quality, consistency, and computational cost evolve as observations accumulate. We introduce HIOcc, a unified benchmark constructed from ScanNet, ScanNet++, and Matterport3D (Dai et al., 2017; Yeshwanth et al., 2023; Chang et al., 2017). As summarized in Fig. 1 and Table 1, it comprises over 2,000 scenes and approximately 300,000 observations, covering about 142,600 m^2. The benchmark combines perspective and panoramic observations under a shared semantic label space, sparse occupancy format, and evaluation framework. Its three tracks connect local prediction, room-level online mapping, and building-level exploration, measuring the quality of individual predictions, their integration over time, and memory construction across connected rooms. Occupancy accuracy and progress AUC measure map quality over exploration; revisit consistency measures agreement when previously observed regions are revisited; and memory footprint and query latency quantify storage and access costs. Together, these evaluations connect the quality of local evidence with the reliability and efficiency of persistent occupancy memory.

Our contributions are summarized as follows:

- We propose GEM-Occ, a Gaussian Evidence Memory framework that allocates memory incrementally and consolidates occupied semantic and free-space evidence through confidence- and visibility-aware causal updates.
- We introduce HIOcc, a unified indoor semantic occupancy benchmark spanning local prediction, room-level online mapping, and building-level exploration.
- Experiments demonstrate improved semantic occupancy accuracy and revisit consistency, together with reduced memory consumption and query latency relative to flat Gaussian memory in building-level mapping.

Table 1: **Occupancy dataset comparison. Mod.**: C/L/R/D = camera/LiDAR/radar/depth. **Coverage Area**: approximate route or scene footprint. For HIOcc, the 12 labels comprise 11 occupied semantic classes and one free-space label.

Dataset	Source Data	Mod.	Surround Views	Scenes	Frames	Classes	Coverage Area
Outdoor / Autonomous Driving							
SemanticKITTI (Behley et al., 2019)	SemanticKITTI, KITTI	C+L	✗	22	4.6K	19	∼3.9 km^2
SSCBench (Li et al., 2024b)	KITTI-360, nuScenes, Waymo	C+L	✓	1859	67K	19	∼43.5 km^2
SurroundOcc (Wei et al., 2023)	nuScenes	C+L	✓	850	40K	17	∼13.2 km^2
OpenOccupancy (Wang et al., 2023)	nuScenes	C+L	✓	850	34K	17	∼13.2 km^2
OpenOcc (Tong et al., 2023)	nuScenes	C+L	✓	850	40K	17	∼13.2 km^2
Occ3D (Tian et al., 2023)	nuScenes, Waymo	C+L	✓	1900	240K	16	∼36.2 km^2
OpenScene (OpenScene Contributors, 2023)	nuPlan	C+L	✓	1.8K	0.4M	2	∼400–500 km^2
Nuplan-Occ (Li et al., 2026b)	nuPlan	C+L	✓	**19K**	**3.6M**	12	**∼400–500 km^2**
Indoor / Embodied AI & Robotics							
MonoScene-NYUv2 (Cao & De Charette, 2022)	NYUv2	C+D	✗	0.5K	1.4K	12	–
Occ-ScanNet (Yu et al., 2024)	ScanNet	C+Depth	✗	681	65.5K	12	∼46.3K m^2
EmbodiedOcc-ScanNet (Wu et al., 2025b)	Occ-ScanNet, ScanNet	C+Depth	✗	681	85.5K	12	∼46.3K m^2
Humanoid Occupancy (Cui et al., 2025)	Self-captured	C+L	✓	200	40K	13	–
HIOcc (Ours)	ScanNet, ScanNet++, Matterport3D	**C+Depth+L**	✓	**2K+**	**∼300K**	**12**	**∼142.6K m^2 total ∼46.3K / ∼31.6K / ∼64.7K m^2**

2 Related Work

2.1 Semantic Occupancy Representations

Voxel-based methods predict occupied regions and semantic categories from depth or RGB (Song et al., 2017; Cao & De Charette, 2022; Yu et al., 2024), using learned lifting and structured features (Li et al., 2023; Zhang et al., 2023; Li et al., 2024a; 2025). In autonomous driving, occupancy models incorporate surround-view observations (Huang et al., 2023; Wei et al., 2023) and temporal reasoning (Ma et al., 2024; Chen et al., 2025; Li et al., 2026a), while self-supervised and generative approaches explore additional formulations for visual scene modeling (Huang et al., 2024a; Zhang et al., 2025a; Li et al., 2026b;c). Building on 3D Gaussian Splatting (Kerbl et al., 2023), Gaussian occupancy methods investigate probabilistic aggregation, geometry guidance, and progressive refinement (Huang et al., 2024b; 2025; Gan et al., 2025; Zhao et al., 2026; Yan & Xu, 2026), as well as language-aligned semantics (Zhou et al., 2026c). Geometry priors are also used for Gaussian occupancy prediction (Qian et al., 2026; Zhou et al., 2026a;b) and occupancy prediction with flexible image inputs (Cao & Vu, 2026). Beyond representations, other work studies native 3D supervision (Boeder et al., 2026) and evidential modeling (Kälble et al., 2025) for occupancy learning.

2.2 Online Occupancy Mapping and Spatial Memory

Online methods combine reconstruction with semantic completion (Wu et al., 2020) or refine global Gaussian memory (Wu et al., 2025b), incorporating geometric and uncertainty-aware updates (Wang et al., 2025a; Zhang et al., 2025b; Guo et al., 2026). Incremental fusion provides another approach: GPOcc fuses Gaussian predictions over time (Zhou et al., 2026a), while FreeOcc combines SLAM with open-vocabulary Gaussian mapping (Jiang et al., 2026). Panoramic systems further extend embodied perception to multimodal sensing (Cui et al., 2025). Related visual geometry models provide geometric priors (Wang et al., 2024a; Leroy et al., 2024; Wang et al., 2025b; Keetha et al., 2026; Li et al., 2026d), while recurrent and streaming states support sequential reconstruction (Wang et al., 2025c; Wu et al., 2025a; Chen et al., 2026). GEM-Occ incrementally builds occupancy memory from semantic Gaussian and free-space evidence, using confidence- and visibility-aware fusion and spatial submaps for memory organization.

2.3 Benchmarks for Embodied Semantic Occupancy

NYUv2, ScanNet, ScanNet++, and Matterport3D (Silberman et al., 2012; Dai et al., 2017; Yeshwanth et al., 2023; Chang et al., 2017) support indoor occupancy and embodied scene understanding benchmarks (Yu et al., 2024; Wu et al., 2020; 2025b; Wang et al., 2024b). Related benchmarks address driving (Wang et al., 2023; Tian et al., 2023; Li et al., 2024b; OpenScene Contributors, 2023), human-aware (Kim et al., 2026), panoramic (Cui et al., 2025; Shi et al., 2026), and open-vocabulary occupancy (Jiang et al., 2026). HIOcc unifies local prediction, room-level online mapping, and building-level mapping within a hierarchical benchmark spanning perspective and panoramic observations. Further discussion is provided in the supplementary material.

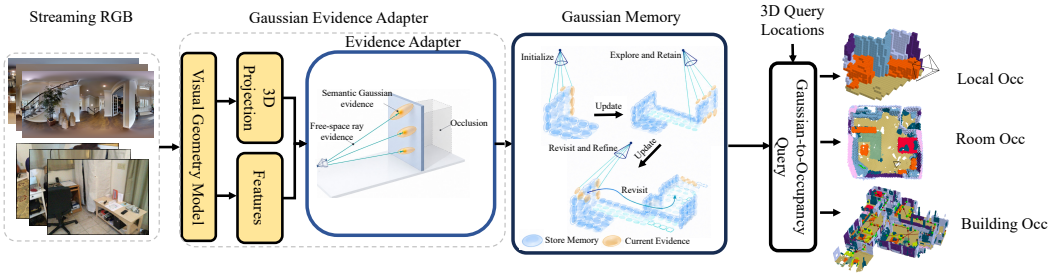

Figure 2: **Overview of GEM-Occ: long-term evidence forms reliable memories.** Streaming perspective and panoramic observations provide geometric predictions and visual features, which the Gaussian evidence adapter converts into world-space semantic Gaussian evidence and free-space rays. Causal memory updates allocate newly observed structures and consolidate evidence through exploration and revisits. Confidence-weighted fusion reinforces supported structures, while free-space evidence revises contradicted occupancy and visibility-aware updates retain occluded memory. Gaussian-to-occupancy queries read out local, room-level, and building-level semantic occupancy from the accumulated memory.

3 GEM-OCC

GEM-Occ builds global semantic occupancy memory by allocating map elements as observations arrive and consolidating occupied and free-space evidence over time. Fig. 2 shows how observation evidence, causal fusion, hierarchical maintenance, and occupancy queries support memory construction across exploration and revisits.

3.1 OBSERVATION-CONDITIONED OCCUPANCY EVIDENCE

Each observation contributes semantic Gaussian evidence and free-space rays for memory updates.

We use a VGGT-based encoder (Wang et al., 2025b) to extract features F_t from each posed observation, represented by a perspective frame or the calibrated sub-views of a panorama. Prediction heads map these features to geometry D_t, semantic logits S_t, and confidence Q_t. Each valid surface prediction yields

$$g_i^t = (\mu_i^t, \Sigma_i^t, \alpha_i^t, p_i^t, \eta_i^t), \tag{1}$$

where μ_i^t is the center transformed into the shared world frame using the supplied calibration and poses, and Σ_i^t is its covariance. The Gaussian evidence adapter predicts opacity α_i^t from $F_t(u)$, while $p_i^t = \mathrm{softmax}(S_t(u))$ and $\eta_i^t = Q_t(u)$ specify semantics and confidence at image location u. Opacity weights local occupancy splatting; confidence weights memory fusion.

Viewing geometry and depth uncertainty determine the spatial support:

$$\Sigma_i^t = R_i^t \mathrm{diag}\left(\sigma_{\perp,i}^2, \sigma_{\perp,i}^2, \sigma_{\parallel,i}^2\right) (R_i^t)^\top. \tag{2}$$

Here, R_i^t aligns the longitudinal axis with the world-frame viewing ray; $\sigma_{\perp,i}$ follows the projected pixel footprint, and $\sigma_{\parallel,i}$ reflects depth uncertainty. Free-space evidence extends along each valid ray before its predicted surface hit, allowing visible contradictions to revise occupancy while preserving occluded memory. Further details on observation processing, covariance scale construction, and parameter settings are provided in the supplementary material.

3.2 INCREMENTAL MEMORY ALLOCATION AND CAUSAL FUSION

The memory $\mathcal{M}_t = (\mathcal{G}_t^{\mathrm{occ}}, \mathcal{R}_t^{\mathrm{free}})$ stores occupied Gaussians and a sparse free-space evidence cache. Each memory Gaussian contains

$$G_j^t = (\mu_j^t, \Sigma_j^t, \ell_j^t, p_j^t, w_j^t, n_j^t), \tag{3}$$

where ℓ_j^t is occupancy log-odds, w_j^t is accumulated evidence weight, and n_j^t counts supporting observations. Incoming evidence is fused with the closest nearby primitive whose squared Mahalanobis

distance, computed using the memory covariance, is below τ_m; unmatched evidence allocates a new primitive. For a matched pair, confidence-weighted fusion updates geometry and semantics:

$$\begin{aligned} \bar{w}_j^t &= \lambda w_j^{t-1} + \eta_i^t, \\ \mu_j^t &= \frac{\lambda w_j^{t-1} \mu_j^{t-1} + \eta_i^t \mu_i^t}{\bar{w}_j^t}, \\ p_j^t &= \frac{\lambda w_j^{t-1} p_j^{t-1} + \eta_i^t p_i^t}{\bar{w}_j^t}. \end{aligned} \tag{4}$$

Here, $\lambda \in [0, 1]$ discounts historical evidence and $w_j^t = \bar{w}_j^t$. Covariance follows a weighted moment merge that accounts for both primitive covariances and their center displacement.

Supporting surfaces and contradictory free-space observations update occupancy:

$$\ell_j^t = \ell_j^{t-1} + \eta_i^t \Delta\ell_{\text{occ}} - \beta_j^t \Delta\ell_{\text{free}}, \tag{5}$$

where $\Delta\ell_{\text{occ}}, \Delta\ell_{\text{free}} > 0$ are the log-odds increments, and $\beta_j^t \in [0, 1]$ measures overlap of the primitive's spatial support with currently observed free-space segments. The occupied term is zero without a surface match; only visible segments before predicted hits contribute to β_j^t and are accumulated in $\mathcal{R}_t^{\text{free}}$. These causal updates reinforce supported structures, suppress contradicted occupancy, and retain occluded memory. Association and fusion details, including parameter settings and free-space overlap computation, are provided in the supplementary material.

3.3 Hierarchical Memory Organization and Maintenance

We organize memory into a local cache, persistent spatial submaps, and a building-level connectivity graph. World-space indexing allocates submaps as new regions are observed and assigns observations by camera pose; revisits update the stored Gaussians through causal fusion. Graph nodes represent submaps, with edges following the provided observation-graph connectivity.

Maintenance merges nearby primitives with overlapping covariance support and compatible semantic distributions:

$$D_{\text{SKL}}(p_j^t, p_k^t) = D_{\text{KL}}(p_j^t \| p_k^t) + D_{\text{KL}}(p_k^t \| p_j^t) < \tau_{\text{sem}}, \tag{6}$$

where τ_{sem} is the semantic compatibility threshold. Merged parameters follow the confidence-weighted fusion rule in Section 3.2. Pruning removes primitives with persistently low occupancy, high semantic entropy, or repeated free-space contradictions. Details of submap organization and memory maintenance are provided in the supplementary material.

3.4 Occupancy Queries and Learning Objectives

For a query location x, let $\mathcal{N}_t(x)$ contain nearby memory Gaussians and $\pi_j^t = \text{sigmoid}(\ell_j^t)$ denote their occupancy probabilities. Gaussian responses yield

$$\begin{aligned} \kappa_j^t(x) &= \exp\left[-\frac{1}{2}(x - \mu_j^t)^\top (\Sigma_j^t)^{-1}(x - \mu_j^t)\right], \\ P_{\text{occ}}^t(x) &= 1 - \prod_{G_j^t \in \mathcal{N}_t(x)} \left(1 - \pi_j^t \kappa_j^t(x)\right). \end{aligned} \tag{7}$$

Semantic predictions use the same spatial and occupancy weights:

$$P_{\text{sem}}^t(x) = \frac{\sum_{G_j^t \in \mathcal{N}_t(x)} \pi_j^t \kappa_j^t(x) p_j^t}{\sum_{G_j^t \in \mathcal{N}_t(x)} \pi_j^t \kappa_j^t(x) + \epsilon}, \tag{8}$$

where $\epsilon > 0$ ensures numerical stability. A query is occupied when $P_{\text{occ}}^t(x) > \tau_{\text{occ}}$, with threshold $\tau_{\text{occ}} \in (0, 1)$ and class $\arg\max_c P_{\text{sem},c}^t(x)$. Remaining queries are free when supported by $\mathcal{R}_t^{\text{free}}$, and unknown otherwise.

Training unrolls memory fusion over multiple observations and supervises memory queries together with auxiliary local predictions obtained by splatting Gaussians with their opacities α_i^t. The visual prediction network and evidence adapter minimize

$$\mathcal{L} = \lambda_{\text{occ}} \mathcal{L}_{\text{occ}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}} + \lambda_{\text{ray}} \mathcal{L}_{\text{ray}}, \tag{9}$$

where the λ coefficients balance the objectives. Occupancy and semantic losses supervise valid occupied/free states and occupied-voxel semantics in both local and memory predictions. The ray loss penalizes occupancy along observed free-space segments in both outputs, while the geometry loss supervises predicted surfaces. Unknown targets are excluded from occupancy supervision. Further details on occupancy readout and supervision of local and fused-memory predictions are provided in the supplementary material.

4 HIOcc and Evaluation Protocol

We construct HIOcc from calibrated observations and annotated geometry in ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and Matterport3D (Chang et al., 2017). Each sample is one posed perspective frame or one panorama with its calibrated sub-views. The shared label space in Table 1 comprises the 11 occupied classes of Occ-ScanNet (Yu et al., 2024) and one free-space label; unknown locations are excluded from evaluation.

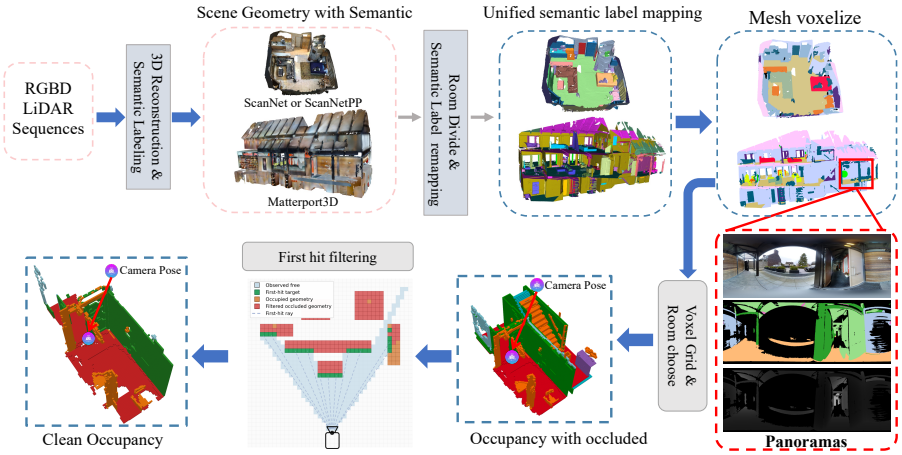

Figure 3: **HIOcc annotation pipeline.** Annotated scene geometry is mapped to shared semantic labels and converted into sparse occupancy targets through viewpoint-dependent cropping and filtering.

As shown in Fig. 3, annotation construction maps source categories to the shared taxonomy, voxelizes scene geometry, and crops and filters targets for each viewpoint. Perspective targets use frustum and depth-consistency filtering, while panoramic targets use first-hit filtering from the shared panorama center. Table 2 evaluates 2D–3D semantic consistency between projected occupied voxels and image-level labels. The full pipeline, including multiview semantic validation, raises semantic-consistency mIoU from 56.5 for mesh voxelization alone to 74.9. Target formats, grid specifications, and filtering details are provided in the supplementary material.

Table 2: **Ablation of HIOcc annotation construction.** Entries report 2D–3D semantic-consistency IoU for the 11 occupied classes. **Complete** denotes a complete-geometry voxel reference, e.g., CompleteScanNet (Wu et al., 2020); **Mesh** denotes the mesh-based semantic voxel source. **Ray-depth** denotes depth-consistency checks for perspective observations and first-hit filtering for panoramic observations. **MV Sem.** denotes multiview semantic validation.

Complete	Mesh	Crop	Frustum	Ray-depth	MV Sem.	ceiling	floor	wall	window	chair	bed	sofa	table	TV	furniture	objects	mIoU
✓	–	–	–	–	–	79.0	88.0	78.5	55.5	68.0	77.0	72.5	63.0	51.5	69.0	63.5	69.6
–	✓	–	–	–	–	72.4	82.1	66.5	39.7	53.9	62.4	57.8	47.8	35.6	52.7	50.6	56.5
–	✓	✓	–	–	–	73.8	83.6	68.1	41.2	55.2	64.0	59.1	49.4	37.1	54.0	52.3	58.0
–	✓	✓	✓	–	–	80.6	87.9	74.6	49.8	62.8	70.7	66.2	57.1	44.3	61.5	58.9	64.9
–	✓	✓	✓	✓	–	86.9	91.2	81.7	58.6	70.5	78.9	74.4	65.8	53.2	69.8	67.4	72.6
–	✓	✓	✓	✓	✓	88.2	92.5	83.4	61.3	73.1	81.0	76.8	68.6	56.4	72.1	70.2	**74.9**

Local prediction and room-level mapping use ScanNet and ScanNet++, with single posed RGB observations and causal perspective sequences, respectively. Building-level mapping uses connected panoramic viewpoints across rooms in Matterport3D. Methods within each track share target grids, semantic labels, valid masks, and observation protocols. We report occupancy IoU and semantic mIoU over the 11 occupied classes, with online mapping additionally evaluated through progress AUC, revisit consistency, memory footprint, and query latency. Progress AUC measures the area under the occupancy-accuracy curve over exploration progress. Revisit consistency measures prediction agreement in previously observed regions upon revisiting.

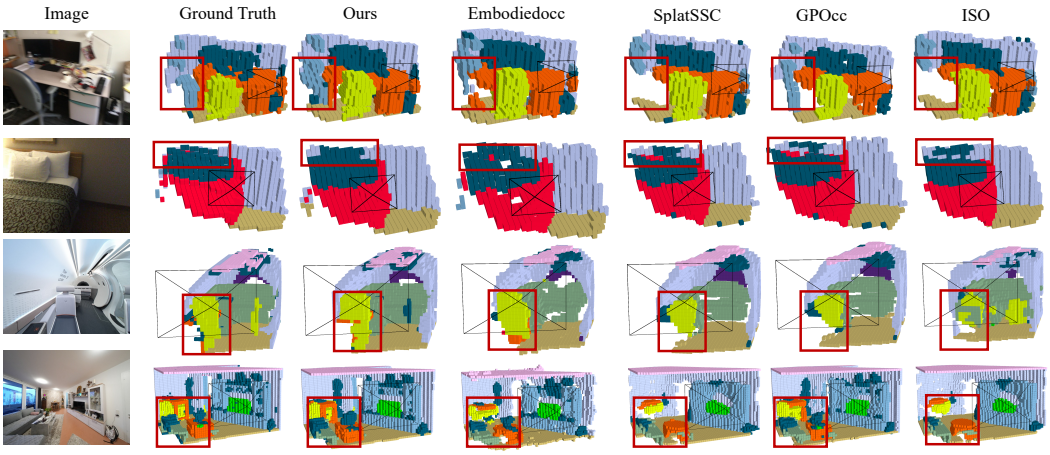

Figure 4: **Qualitative local occupancy prediction.** GEM-Occ produces cleaner semantic structure and fewer spurious occupied regions on HIOcc compared to other methods.

5 EXPERIMENTS

We evaluate the quality of local occupancy evidence, its integration into persistent room-level maps, and the accuracy–efficiency trade-off during building-level exploration. Component ablations examine how evidence representation, free-space constraints, confidence weighting, and memory maintenance contribute to these results.

Experimental Setup. We follow the three evaluation tracks defined in Section 4. Local and room-level experiments use the ScanNet and ScanNet++ perspective-image splits, while building-level experiments use panoramic sequences from Matterport3D. For the building-level setting, we train the evidence encoder and Gaussian evidence adapter on Matterport3D with supervision on fused-memory outputs and auxiliary supervision on local predictions. Methods within each track share semantic labels, target grids, valid masks, and observation protocols. We report occupancy IoU and semantic mIoU, together with progress AUC, revisit consistency, memory footprint, and query latency for online mapping. Local baselines include MonoScene (Cao & De Charette, 2022), ISO (Yu et al., 2024), SplatSSC (Qian et al., 2026), EmbodiedOcc (Wu et al., 2025b), EmbodiedOcc++ (Wang et al., 2025a), and GPOcc (Zhou et al., 2026a). Room-level comparisons additionally include SplicingOcc (Wu et al., 2025b). Further experimental settings and robustness evaluations are provided in the supplementary material.

Local Semantic Occupancy Prediction. GEM-Occ improves both occupancy recovery and semantic prediction from individual observations. As shown in Table 3, it achieves 61.37 IoU and 57.76 mIoU, exceeding GPOcc by 0.68 and 2.68, respectively. The improvement covers all 11 semantic categories, with the largest gain on ceilings (+20.49) and additional gains on chairs (+2.40) and sofas (+1.59). These results indicate that the benefit extends across structural surfaces and object categories, providing more accurate semantic evidence for subsequent map construction. Figure 4 illustrates this behavior: the highlighted regions show more complete structures where several baseline predictions contain missing or fragmented occupancy. The contribution of the evidence representation is further examined through the pointmap and semantic-Gaussian ablations below.

Room-level Online Occupancy Mapping. The accuracy advantage is retained when observations are integrated into persistent room-scale maps. Table 4 reports 56.79 IoU and 46.20 mIoU for GEM-Occ, improving over GPOcc by 3.85 and 1.39, respectively. Compared with EmbodiedOcc, semantic mIoU increases by 5.89. Improvements over GPOcc occur across all 11 categories, including walls (+4.07) and chairs (+2.85), showing that both structural and object-level predictions benefit in the accumulated map. In GEM-Occ, supporting observations refine existing estimates, while observed free space can revise contradicted occupancy and confidence weighting controls the influence of incoming evidence. The free-space and confidence ablations below support the contribution of these updates to room-level accuracy and revisit consistency.

Table 3: **Local semantic occupancy prediction on HIOcc.** We report occupancy IoU and per-class semantic IoU from a single posed RGB observation.

Method	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	TV	furniture	objects	mIoU
MonoScene (Cao & De Charette, 2022)	50.54	74.81	62.33	35.28	35.74	32.07	44.45	39.27	42.66	34.00	32.06	26.64	41.76
ISO (Yu et al., 2024)	52.75	74.91	63.89	37.06	33.67	32.76	43.09	37.65	44.74	34.73	33.42	27.49	42.13
SplatSSC (Qian et al., 2026)	58.15	67.69	68.09	46.21	44.61	38.24	52.50	46.08	47.53	42.69	39.14	34.53	47.94
EmbodiedOcc (Wu et al., 2025b)	53.58	64.33	70.73	38.21	38.26	34.23	46.22	42.91	44.87	34.95	36.12	30.13	43.72
EmbodiedOcc++ (Wang et al., 2025a)	55.62	64.91	71.23	39.18	39.32	34.59	46.81	43.29	45.14	35.62	36.63	30.34	44.28
GPOcc (Zhou et al., 2026a)	60.69	65.96	76.83	52.38	52.08	45.29	58.95	55.43	56.51	53.62	47.97	40.85	55.08
GEM-Occ	**61.37**	**86.45**	**77.31**	**52.92**	**52.46**	**47.69**	**59.77**	**57.02**	**57.40**	**54.64**	**48.37**	**41.37**	**57.76**

Table 4: **Room-level online occupancy mapping on HIOcc.** All methods follow the same causal observation protocol. We report occupancy IoU and per-class semantic IoU after online fusion.

Method	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	TV	furniture	objects	mIoU
SplicingOcc (Wu et al., 2025b)	44.78	53.12	62.34	36.41	33.72	32.12	43.98	36.64	43.71	34.12	33.28	27.43	39.72
EmbodiedOcc (Wu et al., 2025b)	45.12	52.21	62.41	36.74	34.81	32.09	44.82	37.41	44.81	35.86	33.73	28.52	40.31
EmbodiedOcc++ (Wang et al., 2025a)	46.45	53.41	64.43	37.91	33.82	32.89	44.93	38.19	42.03	35.79	33.75	28.82	40.54
GPOcc (Zhou et al., 2026a)	52.94	53.82	66.71	43.96	45.01	36.93	48.98	42.33	45.98	38.91	36.34	33.93	44.81
GEM-Occ	**56.79**	**54.93**	**66.89**	**48.03**	**46.89**	**39.78**	**49.21**	**43.04**	**47.13**	**39.17**	**37.31**	**35.79**	**46.20**

Building-level Occupancy Mapping. Building-level evaluation examines whether accurate and consistent memory can be maintained as exploration extends across connected rooms. The variants in Table 5 share the same Matterport-trained predictor and supervision, allowing comparison of different memory representations and update strategies. Relative to flat Gaussian memory, GEM-Occ raises mIoU from 43.1 to 46.7, progress AUC from 37.2 to 41.5, and revisit consistency from 80.8 to 86.2. The higher progress AUC indicates better accumulated map quality over the course of exploration, while the consistency improvement shows more stable predictions when previously observed regions are revisited. Figure 5 illustrates how the map expands as additional panoramic viewpoints expose connected regions.

These quality improvements accompany lower storage and access costs than flat Gaussian memory. Memory per explored square meter decreases from 1.36 to 0.81 MB/m^2, and dense-query latency decreases from 49.5 to 33.8 ms, corresponding to reductions of 40.4% and 31.7%. Thus, the organized memory provides higher mapping quality with a smaller footprint and lower query cost in this comparison. Post-memory fusion uses slightly less memory and query time, but achieves lower mIoU and revisit consistency. The full model therefore offers a different accuracy–efficiency trade-off, with its gains in persistent-map quality supported by the memory-organization and maintenance ablations below.

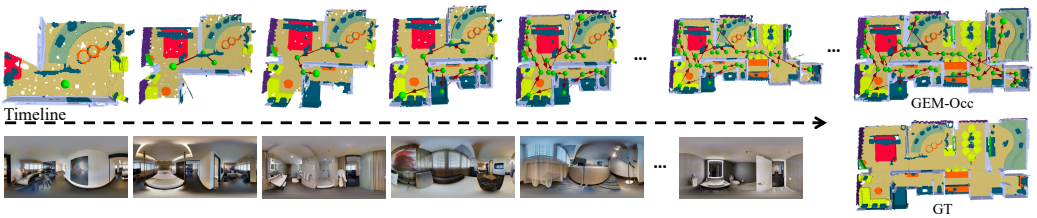

Figure 5: **Streaming building-level (multi-room) mapping.** GEM-Occ incrementally builds semantic occupancy over connected panoramic observations.

Ablation Studies. Table 6 first examines the representation used to construct observation evidence. Replacing Gaussian evidence with pointmap voxelization reduces local mIoU from 57.76 to 53.31 and room-level mIoU from 46.20 to 44.72, while revisit consistency drops from 88.6 to 76.8. The

Table 5: **Building-level occupancy mapping on Matterport3D.** Memory is reported per explored square meter; query latency is measured for dense occupancy queries.

Method	IoU	mIoU	Progress AUC	Revisit Cons.	Mem. MB/m^2	Query Lat. / ms
Per-frame fusion	38.6	31.8	24.7	62.4	0.42	18.3
Pointmap fusion	42.9	35.6	28.9	68.7	0.58	24.6
Post-memory fusion	48.7	41.3	35.8	76.5	0.73	31.2
Flat Gaussian memory	50.4	43.1	37.2	80.8	1.36	49.5
GEM-Occ	53.8	46.7	41.5	86.2	0.81	33.8

simultaneous losses in local prediction and revisit consistency support the use of Gaussian evidence for both observation-level estimation and persistent mapping. Removing semantic Gaussian evidence produces the largest local mIoU decrease among the evaluated variants, to 47.12, and also lowers room-level mIoU to 43.08. This highlights the importance of the semantic evidence representation before and after temporal integration.

Free-space constraints and confidence weighting affect how observations modify the accumulated memory. Without free-space ray evidence, room-level mIoU falls to 42.62 and revisit consistency to 78.9. This behavior is consistent with the role of free-space rays in revising stored occupancy when later observations provide contradictory evidence. Removing confidence-weighted fusion lowers room-level mIoU to 42.83 and revisit consistency to 81.7, while the reported memory footprint remains 0.82 MB/m^2. The quality difference at the same storage cost supports weighting incoming evidence when updating existing estimates.

Memory organization and maintenance determine how much stored information is needed to sustain mapping quality. Removing hierarchical organization increases memory from 0.82 to 1.43 MB/m^2 while reducing room-level mIoU from 46.20 to 43.76. Disabling pruning and merging more than doubles memory usage to 1.78 MB/m^2, yet yields slightly lower room-level mIoU (45.82) and revisit consistency (87.3). The additional stored primitives therefore provide no accuracy benefit in this ablation. These results support consolidation and selective maintenance as a means of retaining a compact memory while preserving useful geometric and semantic evidence.

Table 6: **Ablation of GEM-Occ.** Local metrics are computed from single-frame predictions; room-level metrics are computed after causal fusion. Revisit consistency and memory are measured in the room-level mapping setting, with memory reported per explored square meter.

Setting	Local IoU	Local mIoU	Room IoU	Room mIoU	Revisit Cons.	Mem. MB/m^2
Full GEM-Occ	61.37	57.76	56.79	46.20	88.6	0.82
w/ Pointmap voxelization	57.84	53.31	50.96	44.72	76.8	1.05
w/o Semantic Gaussian evidence	58.63	47.12	52.31	43.08	80.4	0.84
w/o Free-space ray evidence	60.42	51.36	52.74	42.62	78.9	0.80
w/o Confidence-weighted fusion	60.91	52.02	54.18	42.83	81.7	0.82
w/o Hierarchical memory	61.25	52.41	55.02	43.76	83.5	1.43
w/o Pruning/merging	61.31	52.48	56.21	45.82	87.3	1.78

6 Conclusion

We presented GEM-Occ, a Gaussian Evidence Memory framework that turns sequential visual evidence into persistent semantic occupancy memory. The framework constructs occupied semantic Gaussians and free-space rays from each observation, allocates memory as new regions are explored, and consolidates supporting and conflicting evidence through confidence- and visibility-aware causal updates. Hierarchical organization supports continued mapping and occupancy queries across connected indoor spaces. We introduced HIOcc as a unified benchmark for evaluating local prediction, online room mapping, and building-level exploration under shared semantic labels and evaluation protocols. Experiments show improved semantic occupancy accuracy and revisit consistency over the evaluated mapping baselines, together with lower memory consumption and query latency than flat Gaussian memory in building-level mapping. These results demonstrate how evidence accumulated through exploration and revisits supports reliable and scalable spatial memory.

AI USE STATEMENT

We used generative AI tools to assist with implementing code, to improve the clarity and readability of the manuscript, and to suggest revisions to its structure and organization. All research ideas, methods, and conclusions were developed by the authors. The authors bear full responsibility for the content of the manuscript, including any text generated or polished by an LLM, and have ensured that it complies with ethical guidelines and does not involve plagiarism or scientific misconduct.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that our results are reproducible. The main paper and supplementary material describe the architecture, evidence construction, memory fusion, and learning objectives of GEM-Occ. Section 4 and the supplementary material detail HIOcc construction, including semantic label mapping, occupancy target generation, and visibility filtering. Section 5 describes the experimental setup and evaluation protocols, with additional comparisons, runtime measurements, and robustness analyses provided in the supplementary material. HIOcc is constructed from ScanNet, ScanNet++, and Matterport3D. We will release the source code, trained checkpoints, and scripts for dataset construction, training, and evaluation to support independent verification and extension of this work.

REFERENCES

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, 2019.

Simon Boeder, Fabian Gigengack, Simon Roesler, Holger Caesar, and Benjamin Risse. ShelfOcc: Native 3D Supervision beyond LiDAR for Vision-Based Occupancy Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28620–28631, 2026.

Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, 2022.

Anh-Quan Cao and Tuan-Hung Vu. Occany: Generalized unconstrained urban 3d occupancy. In *CVPR*, 2026.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *3DV*, 2017.

Dubing Chen, Huan Zheng, Yucheng Zhou, Xianfei Li, Wenlong Liao, Tao He, Pai Peng, and Jianbing Shen. Semantic causality-aware vision-based 3d occupancy prediction. *ICCV*, 2025.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. In *ICLR*, 2026.

Wei Cui, Haoyu Wang, Wenkang Qin, Yijie Guo, Gang Han, Wen Zhao, Jiahang Cao, Zhang Zhang, Jiaru Zhong, Jingkai Sun, Pihai Sun, Shuai Shi, Botuo Jiang, Jiahao Ma, Jiaxu Wang, Hao Cheng, Zhichao Liu, Yang Wang, Zheng Zhu, Guan Huang, Jian Tang, and Qiang Zhang. Humanoid occupancy: Enabling a generalized multimodal occupancy perception system on humanoid robots. *arXiv preprint arXiv:2507.20217*, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. *ICCV*, 2025.

Yiran Guo, Simone Mentasti, Xiaofeng Jin, Matteo Frosi, and Matteo Matteucci. Sgr-occ: Evolving monocular priors for embodied 3d occupancy prediction via soft-gating lifting and semantic-adaptive geometric refinement. *arXiv preprint arXiv:2603.14076*, 2026.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. *CVPR*, 2023.

Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *CVPR*, 2024a.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. *ECCV*, 2024b.

Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. *CVPR*, 2025.

Zeyu Jiang, Changqing Zhou, Xingxing Zuo, and Changhao Chen. Freeocc: Training-free embodied open-vocabulary occupancy prediction. In *RSS*, 2026.

Jonas Kälble, Sascha Wirges, Maxim Tatarchenko, and Eddy Ilg. EvOcc: Accurate Semantic Occupancy for Automated Driving Using Evidence Theory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27467–27476, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. Mapanything: Universal feed-forward metric 3d reconstruction. In *3DV*, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023.

Junseo Kim, Guido Dumont, Xinyu Gao, Gang Chen, Holger Caesar, and Javier Alonso-Mora. MobileOcc: A Human-Aware Semantic Occupancy Dataset for Mobile Robots. In *Computer Vision – ECCV 2026*, 2026.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024.

Bohan Li, Yasheng Sun, Zhujin Liang, Dalong Du, Zhuanghui Zhang, Xiaofeng Wang, Yunnan Wang, Xin Jin, and Wenjun Zeng. Bridging stereo geometry and bev representation with reliable mutual interaction for semantic scene completion. In *IJCAI*, 2024a.

Bohan Li, Jiajun Deng, Yasheng Sun, Xiaofeng Wang, Xin Jin, and Wenjun Zeng. Hierarchical context alignment with disentangled geometric and temporal modeling for semantic occupancy prediction. *TPAMI*, 2026a.

Bohan Li, Xin Jin, Hu Zhu, Hongsi Liu, Ruikai Li, Jiazhe Guo, Kaiwen Cai, Chao Ma, Yueming Jin, Hao Zhao, Xiaokang Yang, and Wenjun Zeng. Scaling up occupancy-centric driving scene generation: Dataset and method. *TPAMI*, 2026b.

Bohan Li, Zhuang Ma, Dalong Du, Baorui Peng, Zhujin Liang, Zhenqiang Liu, Xianda Guo, Zheng Zhu, Chao Ma, Yueming Jin, Xin Jin, Hao Zhao, and Wenjun Zeng. Omninwm: Omniscient driving navigation world models. In *ECCV*, 2026c.

Bohan Li, Shuojue Yang, Baorui Peng, Xianda Guo, Erli Zhang, Youqi Tao, Junfeng Duan, Daguang Xu, Qi Dou, Xin Jin, et al. From articulated kinematics to routed visual control for action-conditioned surgical video generation. *arXiv preprint arXiv:2605.08712*, 2026d.

Jianing Li, Ming Lu, Juntao Liu, Hao Wang, Chenyang Gu, Wenzhao Zheng, Li Du, and Shanghang Zhang. Sliceocc: Indoor 3d semantic occupancy prediction with vertical slice representation. In *ICRA*, 2025.

Yiming Li, Zhiding Yu, Christopher Bongsoo Choy, Chaowei Xiao, José Manuel Álvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. *CVPR*, 2023.

Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, Yue Wang, Hang Zhao, Zhiding Yu, and Chen Feng. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *IROS*, 2024b.

Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao Gu, Rui Ai, and Hesheng Wang. Cam4DOcc: Benchmark for Camera-Only 4D Occupancy Forecasting in Autonomous Driving Applications. In *CVPR*, 2024.

OpenScene Contributors. Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. `https://github.com/OpenDriveLab/OpenScene`, 2023.

Rui Qian, Haozhi Cao, Tianchen Deng, Shenghai Yuan, and Lihua Xie. Splatssc: Decoupled depth-guided gaussian splatting for semantic scene completion. In *AAAI*, 2026.

Hao Shi, Ze Wang, Shangwei Guo, Mengfei Duan, Song Wang, Teng Chen, Kailun Yang, Lin Wang, and Kaiwei Wang. OneOcc: Semantic Occupancy Prediction for Legged Robots with a Single Panoramic Camera. In *CVPR*, 2026.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. In *NeurIPS Datasets and Benchmarks Track*, 2023.

Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, and Hongyang Li. Scene as occupancy. In *ICCV*, 2023.

Hao Wang, Xiaobao Wei, Xiaoan Zhang, Jianing Li, Chengyu Bai, Ying Li, Ming Lu, Wenzhao Zheng, and Shanghang Zhang. Embodiedocc++: Boosting embodied 3d occupancy prediction with plane regularization and uncertainty sampler. In *ACM MM*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024a.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024b.

Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *ICCV*, 2023.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, 2023.

Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scfusion: Real-time incremental scene reconstruction with semantic completion. In *3DV*, 2020.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. In *NeurIPS*, 2025a.

Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding. In *ICCV*, 2025b.

Chi Yan and Dan Xu. Progressive gaussian transformer with anisotropy-aware sampling for open vocabulary occupancy prediction. In *ICLR*, 2026.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023.

Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In *ECCV*, 2024.

Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Advancing 3d occupancy prediction in lidar-free environments. *IEEE Transactions on Image Processing*, 2025a.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *ICCV*, 2023.

Zhang Zhang, Qiang Zhang, Wei Cui, Shuai Shi, Yijie Guo, Gang Han, Wen Zhao, Hengle Ren, Renjing Xu, and Jian Tang. Roboocc: Enhancing the geometric and semantic scene understanding for robots. *arXiv preprint arXiv:2504.14604*, 2025b.

Lingjun Zhao, Sizhe Wei, James Hays, and Lu Gan. Gaussianformer3d: Multi-modal gaussian-based semantic occupancy prediction with 3d deformable attention. In *ICRA*, 2026.

Changqing Zhou, Yueru Luo, and Changhao Chen. Generalizing visual geometry priors to sparse gaussian occupancy prediction. In *CVPR*, 2026a.

Changqing Zhou, Yueru Luo, Yulan Guo, Bing Wang, Jie Qin, and Changhao Chen. GPOcc++: Unified Sparse Gaussian Occupancy Prediction with Visual Geometry Priors. *arXiv 2607.13481*, 2026b.

Changqing Zhou, Yueru Luo, Han Zhang, Zeyu Jiang, and Changhao Chen. Monocular open vocabulary occupancy prediction for indoor scenes. In *CVPR*, pp. 21627–21637, 2026c.

A Additional Experiments

A.1 External Baseline Comparisons

Table 7 reports adapted baselines under the HIOcc evaluation protocol across local, room-level, and building-level tasks. In building-level mapping, GEM-Occ achieves 53.80 IoU and 46.70 mIoU, exceeding GPOcc by 3.55 and 2.89 points, respectively. These comparisons complement the main-text building-level variants, which use the same Matterport-trained predictor and supervision to examine different memory representations and update strategies. LegoOcc and FreeOcc are evaluated on the same 11 occupied classes.

Table 7: **Additional baseline comparisons on HIOcc.** † denotes adaptation to the shared evaluation protocol. Each task reports occupancy IoU and semantic mIoU over the 11 occupied classes; – indicates an unreported result.

Method	Local		Room		Building	
	IoU	mIoU	IoU	mIoU	IoU	mIoU
EmbodiedOcc† Wu et al. (2025b)	53.58	43.72	45.12	40.31	44.92	40.73
GPOcc† Zhou et al. (2026a)	60.69	55.08	52.94	44.81	50.25	43.81
LegoOcc† Zhou et al. (2026c)	59.33	20.79	–	–	–	–
FreeOcc† Jiang et al. (2026)	–	–	30.21	13.85	28.92	12.56
GEM-Occ	**61.37**	**57.76**	**56.79**	**46.20**	**53.80**	**46.70**

A.2 Runtime and Memory Accounting

Table 8 reports throughput for evidence extraction and the complete online pipeline, measured on the same hardware. Evidence extraction runs at 8.7 FPS before causal fusion. The full pipeline runs at 5.2 FPS and includes evidence extraction, causal fusion, submap and graph maintenance, and occupancy querying.

Table 8: **Runtime of GEM-Occ.** Throughput is measured on the same hardware for the two processing scopes.

Processing scope	Throughput (FPS)
Evidence extraction	8.7
Complete online pipeline	5.2

Memory footprint is reported in MB per explored square meter. The main-text ablation measures memory and revisit consistency in the room-level setting, while the building-level comparison measures them on panoramic sequences across connected rooms. The full model uses 0.82 MB/m^2 in the room-level ablation and 0.81 MB/m^2 in the building-level comparison. Against flat Gaussian memory in the building-level setting, memory decreases from 1.36 to 0.81 MB/m^2 and dense-query latency from 49.5 to 33.8 ms, reductions of 40.4% and 31.7%, respectively. Dense-query latency measures occupancy readout and is distinct from the time required to process a new observation through the complete pipeline.

A.3 Robustness to Pose and Depth Perturbations

We evaluate sensitivity to pose and geometric evidence using controlled translation, rotation, and relative depth perturbations in the building-level setting. Table 9 reports occupancy IoU and semantic mIoU for each perturbation level. Increasing translation from 0.1 to 0.2 m reduces mIoU from 40.33 to 34.54; increasing yaw perturbation from 5° to 10° reduces it from 39.25 to 31.57. Relative depth noise at $\sigma = 5\%$ and 10% yields 42.78 and 31.16 mIoU, respectively, quantifying the effect of geometric uncertainty on the accumulated memory. The degradation shows that map quality remains sensitive to alignment and surface prediction errors, even with confidence-weighted fusion. These tests characterize input sensitivity in the posed-mapping setting; they do not evaluate trajectory estimation or SLAM drift correction.

Table 9: **Building-level robustness under pose and depth perturbations.**

Setting	IoU	mIoU
Default	**53.80**	**46.70**
Translation 0.1 m (x-axis)	50.14	40.33
Translation 0.2 m (x-axis)	45.76	34.54
Rotation 5° (yaw)	47.29	39.25
Rotation 10° (yaw)	42.51	31.57
Relative depth noise ($\sigma = 5\%$)	44.64	42.78
Relative depth noise ($\sigma = 10\%$)	39.22	31.16

Scope and limitations. The current evaluation concerns static indoor scenes with supplied camera poses and a fixed semantic taxonomy. Building-level experiments train the evidence predictor on Matterport3D; they do not establish zero-shot transfer from perspective-image training to panoramic observations. Dynamic-scene mapping and downstream navigation or manipulation remain outside the evaluated setting.

A.4 Additional Mapping Visualizations

Figure 6 presents two additional building-level mapping sequences on Matterport3D. As panoramic observations arrive, the accumulated map expands across connected regions while retaining previously mapped structures. The final prediction in the upper sequence captures the major room layout and semantic regions visible in the ground-truth map.

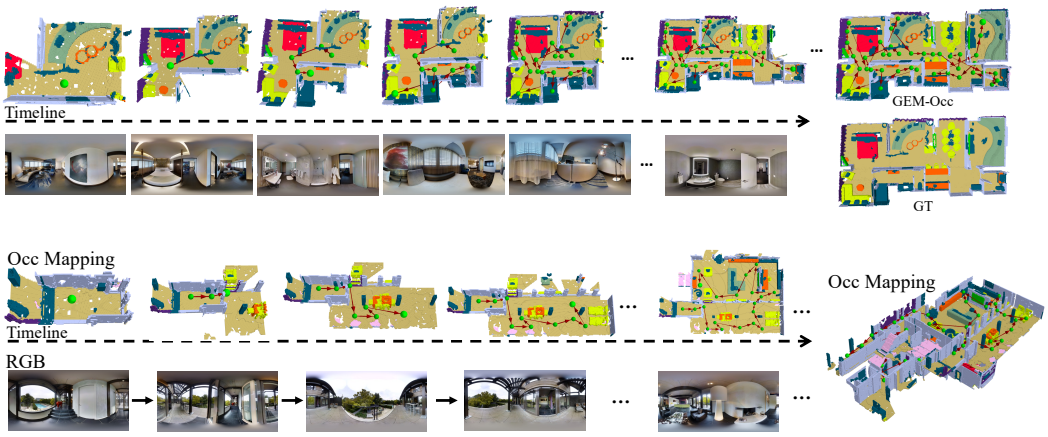

Figure 6: **Additional building-level mapping results.** Two sequences show the evolution of GEM-Occ's semantic occupancy memory alongside the incoming panoramic RGB observations, ordered from left to right. The upper sequence includes the final prediction and ground-truth map on the right; the lower sequence ends with a larger view of the accumulated map.

B GEM-Occ Method Details

B.1 Input and Mapping Protocol

GEM-Occ receives posed RGB observations with camera calibration. A time step corresponds to a perspective frame or a panorama with its calibrated sub-views, whose evidence is expressed in a shared world frame. Depth and semantic annotations provide supervision and benchmark targets; RGB-only inference uses predicted geometry. Camera poses are supplied, and memory updates use only observations available up to the current time step.

B.2 Observation Evidence and Ray Support

The adapted VGGT Wang et al. (2025b) encoder extracts streaming visual features from perspective and panoramic observations. Prediction heads produce local geometry D_t, semantic logits S_t, and confidence Q_t; supplied calibration and poses express the predicted surfaces and viewing rays in a shared world coordinate system. For a valid image location u, the Gaussian evidence adapter predicts occupancy opacity α_i^t from $F_t(u)$, while $p_i^t = \mathrm{softmax}(S_t(u))$ and $\eta_i^t = Q_t(u)$ provide the semantic distribution and incoming evidence confidence. Opacity controls the contribution to local occupancy splatting, whereas confidence determines the weight of the observation during memory fusion.

The ray-aligned covariance associates each surface prediction with volumetric support determined by its viewing geometry and depth uncertainty. Its two transverse axes follow the projected pixel footprint, and its longitudinal axis follows the world-frame viewing ray, with scale reflecting uncertainty in predicted surface depth. Free-space evidence is recorded only on visible ray segments before the predicted surface hit. These segments identify where the current observation can contradict stored occupancy; regions beyond the hit contribute no free-space evidence and retain their existing memory in the absence of other observations.

B.3 Spatial Association and Covariance Fusion

Incoming evidence is associated with memory through a local spatial search and a squared Mahalanobis-distance gate. For an evidence primitive g_i^t and a candidate G_j^{t-1}, the distance is

$$d_M(g_i^t, G_j^{t-1}) = (\mu_i^t - \mu_j^{t-1})^\top (\Sigma_j^{t-1})^{-1} (\mu_i^t - \mu_j^{t-1}). \tag{10}$$

The closest candidate satisfying $d_M < \tau_m$ receives the incoming evidence; if none satisfies the gate, the evidence allocates a new memory primitive. An unmatched evidence primitive initializes a new memory element with its predicted center, covariance, and semantic distribution. Allocation therefore follows the spatial coverage of observed structures, while repeated observations consolidate existing map elements.

For a matched pair, the main-text confidence-weighted rule updates the center and semantic distribution with total weight $\bar{w}_j^t = \lambda w_j^{t-1} + \eta_i^t$. Covariance is updated by a weighted moment merge:

$$\begin{aligned} \Sigma_j^t = \frac{1}{\bar{w}_j^t} \big[& \lambda w_j^{t-1} \big(\Sigma_j^{t-1} + \delta_j^{t-1} (\delta_j^{t-1})^\top \big) \\ & + \eta_i^t \big(\Sigma_i^t + \delta_i^t (\delta_i^t)^\top \big) \big], \end{aligned} \tag{11}$$

where $\delta_j^{t-1} = \mu_j^{t-1} - \mu_j^t$ and $\delta_i^t = \mu_i^t - \mu_j^t$. The displacement terms account for the separation of the original centers around the fused center, preserving the spatial spread of the combined evidence.

When multiple evidence primitives from the same observation are associated with one memory element, their geometric and semantic estimates are first aggregated using confidence-weighted moments, including the covariance contributions from center displacement. The summed confidence serves as the aggregate weight. The single-pair equations above then fuse this aggregate with the stored memory primitive. The historical discount and free-space correction are each applied once per memory element per observation.

Occupancy updates combine the positive increment $\eta_i^t \Delta\ell_{\mathrm{occ}}$ from matched surfaces with the negative increment $\beta_j^t \Delta\ell_{\mathrm{free}}$ from observed free space. The occupied increment is zero when there is no surface match. Let $\mathcal{S}_j^t$ be a finite set of sample locations within the primitive's spatial support, and let $\mathcal{F}_t$ denote the discretized free-space region supported by the current observation. We compute the overlap as

$$\beta_j^t = \frac{\sum_{x \in \mathcal{S}_j^t} \kappa_j^t(x)\, \mathbf{1}[x \in \mathcal{F}_t]}{\sum_{x \in \mathcal{S}_j^t} \kappa_j^t(x)}, \tag{12}$$

where $\kappa_j^t(x)$ is the Gaussian spatial response defined in the main text. This normalized overlap lies in $[0, 1]$; we set $\beta_j^t = 0$ when the sample set is empty. Only segments before predicted surface hits contribute negative evidence, so occlusion alone does not reduce stored occupancy. An existing primitive can receive a free-space correction even without a surface match in the current observation.

If neither supporting surface evidence nor contradictory free-space evidence is available, its occupancy log-odds remains unchanged. The correction uses only the current observation's free-space evidence. The same segments accumulate in the sparse cache $\mathcal{R}_t^{\text{free}}$ for later queries.

B.4 Submap Organization and Memory Maintenance

World-space indexing creates spatial submaps as new regions are observed and assigns observations using their camera poses. Each submap stores Gaussian memory in the shared world frame, allowing a revisit to update the existing primitives through the same association and fusion rules. The room-level submaps are spatial partitions defined by the index; Matterport3D subroom annotations are used only for benchmark construction. Graph nodes represent submaps, and edges follow the provided observation-graph connectivity. The observation graph supplies the sequence, poses, and adjacency for this organization. The graph records connections among mapped regions, while the spatial index supports access to their Gaussian memory for updates and queries.

Maintenance operates periodically within active submaps. Merge candidates have nearby centers, overlapping covariance support, and compatible semantic predictions under the symmetric KL-divergence criterion in the main text. Their parameters are combined using confidence-weighted fusion. Pruning removes primitives whose occupancy confidence remains low after repeated observations, whose semantic entropy stays high, or that are repeatedly contradicted by free-space evidence. Together, these operations consolidate duplicate geometry and remove poorly supported primitives as evidence accumulates.

B.5 Occupancy Readout and Supervision

Occupancy queries can be evaluated at arbitrary 3D locations or sampled on the voxel grid required by each evaluation setting. Each nearby Gaussian contributes according to its spatial response and the occupancy probability obtained from its stored log-odds. The same weights aggregate semantic predictions. Using label 0 for free space and $\perp$ for unknown space, the readout is

$$\hat{y}_t(x) = \begin{cases} \arg\max_c P^t_{\text{sem},c}(x), & P^t_{\text{occ}}(x) > \tau_{\text{occ}}, \\ 0, & P^t_{\text{occ}}(x) \le \tau_{\text{occ}} \text{ and } x \in \mathcal{R}_t^{\text{free}}, \\ \perp, & \text{otherwise.} \end{cases} \tag{13}$$

Occupied evidence takes precedence when Gaussian support and cached free-space support overlap. Locations without occupied or free-space support remain unknown.

Training unrolls memory fusion over multiple observations and supervises fused-memory queries using the view-conditioned targets in HIOcc. Local predictions are formed by splatting the incoming evidence Gaussians with their predicted opacities α_i^t. Auxiliary supervision on these outputs supplies a learning signal for opacity alongside the supervision on the accumulated memory. The occupancy, semantic, and ray objectives apply to both local and fused-memory predictions. Occupancy supervision covers valid occupied and free target locations, semantic supervision covers occupied target voxels, and ray supervision penalizes occupancy along observed free-space segments. The geometry objective supervises predicted surfaces using the available geometric targets. Unknown target locations are excluded from occupancy supervision. The visual prediction network and evidence adapter are learned; the fusion and query rules introduce no additional learned parameters. Supervision is applied to occupancy queries and predicted geometry, without Gaussian-level annotations.

B.6 Training and Evaluation Setup

Local and room-level experiments use the ScanNet and ScanNet++ perspective-image subsets of HIOcc. Building-level experiments use the Matterport3D panoramic subset, with observations represented by panoramas and their calibrated sub-views. For this setting, the evidence encoder and Gaussian evidence adapter are trained on Matterport3D using the shared memory representation and causal fusion rule. Methods within each task are evaluated with the same semantic labels, target grids, and valid masks; online evaluation follows the same causal observation protocol.

C HIOcc Annotation and Target Details

C.1 Viewpoint Samples and Sparse Targets

HIOcc combines calibrated observations and annotated scene geometry from ScanNet Dai et al. (2017), ScanNet++ Yeshwanth et al. (2023), and Matterport3D Chang et al. (2017) within a shared occupancy representation. Perspective observations cover RGB-D scans and higher-fidelity reconstructions, while panoramic observations connect rooms and floors within buildings. A sample is one viewpoint: a posed perspective frame or a panorama with its calibrated sub-views. Depth maps, semantic maps, auxiliary views, and targets at different resolutions support annotation and supervision for that viewpoint and are counted as part of the same sample.

Targets store sparse occupied-voxel entries $[i, j, k, l]$, where (i, j, k) are voxel indices and l is the semantic label, together with grid origin, voxel size, and grid shape. The 11 occupied categories follow Occ-ScanNet Yu et al. (2024): *ceiling*, *floor*, *wall*, *window*, *chair*, *bed*, *sofa*, *table*, *TV*, *furniture*, and *objects*. Together with the free-space label, they form the 12 labels reported for HIOcc. The sparse files explicitly store occupied semantic labels; observation geometry and validity masks determine free and unknown states during supervision and evaluation. Free space is supported by observed empty ray segments, while unknown locations fall outside valid target support and are masked out. Semantic mIoU averages the 11 occupied categories, excluding free space. An absent occupied entry does not by itself establish free space.

C.2 Perspective and Panoramic Target Construction

Annotation construction first maps source categories to the shared taxonomy and voxelizes annotated scene geometry in world coordinates. For each viewpoint, a local metric volume is cropped and filtered according to its observation geometry. For perspective targets, voxel centers are transformed into the camera frame, filtered by the image frustum, and checked for depth consistency when depth is available. The perspective target hierarchy provides three resolutions with a common physical extent (Table 10).

Panoramic targets use a $120 \times 120 \times 60$ grid with $0.05\,\text{m}$ voxels, covering $6 \times 6 \times 3\,\text{m}$ around the shared panorama center. For targets covered by calibrated sub-views, first-hit filtering casts rays from this center and uses ceiling, floor, wall, and window voxels as structural occluders. The filtering retains a small region around the panorama center and a finite thickness of the first-hit structures, then removes occupancy occluded beyond that retained support. This produces targets with volumetric structural support for panoramic observations.

Table 10: **Target grid specifications.** Perspective resolutions describe alternative targets for the same observation.

Target	Grid shape	Voxel size (m)	Extent (m)
Perspective	$240 \times 240 \times 144$	0.02	$4.8 \times 4.8 \times 2.88$
Perspective	$120 \times 120 \times 72$	0.04	$4.8 \times 4.8 \times 2.88$
Perspective	$60 \times 60 \times 36$	0.08	$4.8 \times 4.8 \times 2.88$
Panoramic	$120 \times 120 \times 60$	0.05	$6 \times 6 \times 3$

C.3 Annotation Validation

We validate targets through 2D–3D semantic consistency by projecting occupied voxels into calibrated views and comparing their labels with image-level semantic evidence. The main-text ablation reports per-class consistency after mesh voxelization, local cropping, view filtering, depth or first-hit filtering, and multiview semantic validation. The full construction raises semantic-consistency mIoU from 56.5 to 74.9 across the 11 occupied categories. These scores measure agreement between constructed targets and image-level semantic evidence.

Figure 7 provides qualitative comparisons with Occ-ScanNet and additional annotation examples from ScanNet++ and Matterport3D. In the highlighted ScanNet regions, HIOcc preserves object structures that are incomplete or missing in the Occ-ScanNet targets, improving correspondence with

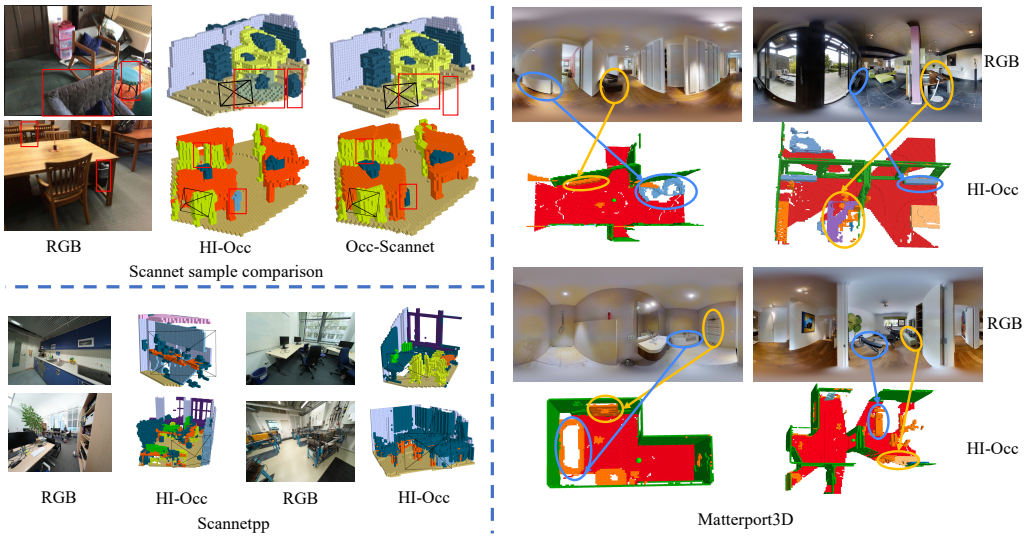

Figure 7: **Qualitative evaluation of HIOcc annotations.** Top left: comparison with Occ-ScanNet on ScanNet, with red boxes highlighting differences in annotated object structure. Bottom left: additional ScanNet++ examples. Right: Matterport3D panoramic observations and occupancy targets, with colored circles and arrows indicating corresponding structures.

the RGB observations. The remaining examples show how the annotations capture visible furniture and room structures under perspective and panoramic observation settings.

C.4 Evaluation Protocol and Metrics

Local prediction uses a single posed RGB observation, room-level mapping uses causal perspective sequences, and building-level mapping follows connected panoramic viewpoints. All comparisons within a track use the same target grids, labels, and valid evaluation masks.

Occupancy IoU measures overlap between predicted and target occupied voxels on valid evaluation support. Semantic mIoU averages the class-wise intersection over union for the $C = 11$ occupied categories:

$$\mathrm{IoU}_{\mathrm{occ}} = \frac{\mathrm{TP}_{\mathrm{occ}}}{\mathrm{TP}_{\mathrm{occ}} + \mathrm{FP}_{\mathrm{occ}} + \mathrm{FN}_{\mathrm{occ}}}, \qquad \mathrm{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{14}$$

Free space is excluded from the semantic class average, and unknown targets are excluded by the evaluation mask.

Progress AUC summarizes the area under the occupancy-accuracy curve over exploration progress. Revisit consistency measures prediction agreement in previously observed regions when the agent returns to them. They complement final-map accuracy by measuring quality throughout exploration and stability upon revisiting.

D Extended Related Work

D.1 Semantic Occupancy Representations

Semantic occupancy prediction jointly estimates the geometry and semantic categories of 3D space. Early semantic scene completion methods infer volumetric structure from depth observations, while monocular approaches recover occupancy and semantics from RGB images Song et al. (2017); Cao & De Charette (2022); Yu et al. (2024). Camera-based methods develop image-to-volume lifting, sparse voxel queries, and structured spatial features Li et al. (2023); Zhang et al. (2023); Li et al. (2024a; 2025). In autonomous driving, surround-view and tri-perspective representations support scene-wide

prediction Huang et al. (2023); Wei et al. (2023), with temporal modeling, self-supervision, and generative formulations extending occupancy to sequential scene understanding Ma et al. (2024); Chen et al. (2025); Huang et al. (2024a); Zhang et al. (2025a); Li et al. (2026a;b;c). OccAny further uses visual geometry models to predict urban occupancy from monocular, sequential, or surround-view images Cao & Vu (2026).

3D Gaussian Splatting Kerbl et al. (2023) motivates continuous occupancy representations built from spatial primitives with geometry and semantic attributes. GaussianFormer Huang et al. (2024b) represents scenes using sparse semantic Gaussians and obtains voxel predictions by aggregating nearby primitives through Gaussian-to-voxel splatting. GaussianFormer-2 Huang et al. (2025) introduces probabilistic Gaussian superposition and distribution-based initialization to improve the allocation and aggregation of occupied-space primitives. GaussianOcc Gan et al. (2025) exploits Gaussian splatting for self-supervised occupancy learning, while GaussianFormer3D Zhao et al. (2026) incorporates LiDAR geometry priors and camera features through 3D deformable attention.

For indoor scenes, SplatSSC Qian et al. (2026) combines depth-guided Gaussian initialization with decoupled geometric and semantic aggregation. GPOcc Zhou et al. (2026a) extends surface predictions along camera rays to construct volumetric Gaussian samples from visual geometry priors. LegoOcc Zhou et al. (2026c) couples occupancy geometry with language-aligned semantic embeddings through language-embedded Gaussians, enabling monocular open-vocabulary occupancy prediction. In GEM-Occ, Gaussian primitives encode occupied semantic evidence, with centers and covariances determined by predicted geometry and its uncertainty. We pair these primitives with explicit free-space ray evidence and consolidate both in an incrementally constructed global memory.

D.2 Online Occupancy Mapping and Spatial Memory

Online scene understanding maintains spatial knowledge as observations arrive during exploration. SCFusion Wu et al. (2020) combines incremental reconstruction with semantic scene completion from depth sequences. EmbodiedOcc Wu et al. (2025b) maintains an explicit global Gaussian memory, initializes Gaussians uniformly across the scene, and refines those inside the current camera frustum. EmbodiedOcc++ Wang et al. (2025a) introduces plane regularization and uncertainty-based selection of Gaussians for refinement. RoboOcc Zhang et al. (2025b) improves Gaussian interactions through opacity-guided and geometry-aware encoders, while SGR-OCC Guo et al. (2026) combines uncertainty-aware feature lifting, ray-constrained refinement, and progressive training for temporal fusion. Humanoid Occupancy Cui et al. (2025) extends embodied occupancy perception to panoramic multimodal sensing on humanoid platforms.

Incremental fusion provides another route to persistent occupancy mapping. GPOcc Zhou et al. (2026a) fuses per-frame Gaussians into a global representation through a training-free update strategy. FreeOcc Jiang et al. (2026) integrates SLAM, incremental Gaussian mapping, and semantic features from pretrained vision-language models to construct open-vocabulary occupancy maps from monocular or RGB-D sequences. These methods connect frame-level geometric predictions with global maps through spatial association and temporal integration.

GEM-Occ develops a global semantic occupancy memory with on-demand allocation and explicit occupied and free-space evidence. Local geometric predictions determine the position and spatial support of incoming semantic Gaussians, which are associated with existing memory elements or inserted as new elements. Confidence-weighted fusion consolidates supporting observations, while free-space rays suppress contradicted occupancy and visibility-aware updates preserve occluded structure. Spatial submaps and a building-level connectivity graph organize the accumulated evidence for continued exploration and occupancy queries.

D.3 Visual Geometry Models as Observation Evidence

Streaming visual geometry models also accumulate information through persistent reconstruction states. Image-based geometry estimators provide depth and pointmap predictions from visual observations Wang et al. (2024a); Leroy et al. (2024); Wang et al. (2025b); Keetha et al. (2026). For streaming reconstruction, CUT3R Wang et al. (2025c) recurrently updates a latent state, Point3R Wu et al. (2025a) associates an explicit pointer memory with 3D locations, and TTT3R Chen et al. (2026) adapts state updates according to the alignment between historical memory and incoming

observations. Their geometric outputs provide surface positions and temporal context for constructing spatial maps.

Geometric predictions and reconstruction states provide complementary information for evidence construction. Predicted surfaces determine where an observation supports occupied structure, and the viewing rays identify observed free space before each surface hit. In GEM-Occ, surface geometry and uncertainty determine the centers and covariances of incoming semantic Gaussians, while image features support semantic and opacity prediction. Causal fusion accumulates this evidence in memory elements that retain occupancy log-odds, semantic estimates, evidence weights, and observation counts. This connects visual geometry estimation with occupancy memory that can be updated and queried throughout exploration.

D.4 Benchmarks for Embodied Semantic Occupancy

Semantic occupancy benchmarks define the observations, spatial targets, and evaluation protocols used to measure 3D scene understanding. NYUv2, ScanNet, ScanNet++, and Matterport3D provide RGB-D observations and semantic scene annotations across different indoor capture settings Silberman et al. (2012); Dai et al. (2017); Yeshwanth et al. (2023); Chang et al. (2017). ScanNet provides posed perspective frames, reconstructed surfaces, and semantic annotations that support local and room-level occupancy evaluation. ScanNet++ supplies higher-fidelity reconstructions and visual observations, while Matterport3D supplies panoramic observations across connected rooms and floors. These capture settings support evaluation of both local scene prediction and memory construction during exploration. In autonomous driving, OpenOccupancy, Occ3D, SSCBench, and OpenScene establish large-scale evaluation of surrounding occupancy and semantic scene completion Wang et al. (2023); Tian et al. (2023); Li et al. (2024b); OpenScene Contributors (2023).

Indoor occupancy benchmarks develop complementary prediction and mapping settings. Occ-ScanNet evaluates local semantic occupancy from monocular observations, while EmbodiedOcc-ScanNet organizes posed observation sequences for room-level online occupancy prediction Yu et al. (2024); Wu et al. (2025b). CompleteScanNet supplies completed scene geometry for semantic scene completion Wu et al. (2020). EmbodiedScan broadens indoor evaluation to multimodal egocentric perception with dense semantic occupancy annotations Wang et al. (2024b). Humanoid Occupancy introduces panoramic multimodal occupancy data for humanoid robots Cui et al. (2025), and ReplicaOcc, introduced with FreeOcc Jiang et al. (2026), supports indoor open-vocabulary occupancy evaluation.

HIOcc establishes a unified benchmark for evaluating semantic occupancy memory from local observations to connected buildings. Its three tracks cover local prediction, causal room-level mapping, and building-level exploration, with shared semantic labels and occupancy-state definitions. View-conditioned annotations and validity masks support consistent evaluation across perspective frames and pano-centric observation groups. Semantic mIoU is computed over the 11 occupied classes inherited from Occ-ScanNet, while free space is evaluated as a separate occupancy state. This common framework connects occupancy accuracy with revisit consistency, memory footprint, and query latency, enabling assessment of how accumulated observations support reliable and scalable spatial memory.