2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MemOcc: Hierarchical Memory for Indoor Continuous Occupancy Mapping

Yirong Yang[1], Yuxin Lin[1], Longteng Guo[2*], Li Song[3], Qunbo Wang[4*], Ming-Ming Yu[1], Wenjun Wu[1], Jing Liu[2,5]

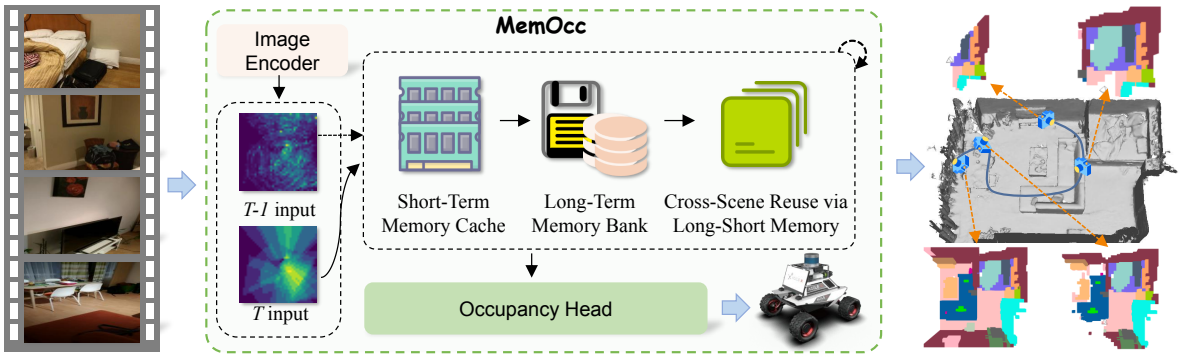

Fig. 1: We propose a memory-augmented framework for indoor embodied 3D perception that couples short-term selective read–write aggregation with long-term scene-prior retrieval, enabling robots to robustly accumulate reliable voxel evidence over long-horizon multi-step observations and quickly adapt to new scenes.

***Abstract*— Indoor 3D occupancy mapping, crucial for robotic perception, struggles with occlusions and reappearing surfaces in continuous observations. Existing methods either fuse frames without discernment, causing occlusion-induced errors to persist and contaminate global representations, or recompute scenes from scratch, sacrificing efficiency and stability. To address these challenges, we propose MemOcc, a novel memory-augmented framework for continuous occupancy mapping using read–write–retrieve operations. MemOcc employs a hierarchical memory design with cooperative short- and long-term tiers. Its Short-Term Memory Cache module uses visibility-gated writes and confidence maps to stabilize voxel predictions and filter occlusion noise, while the Long-Term Memory Bank stores scene priors for rapid retrieval, accelerating convergence in revisited regions. As a plug-and-play module, MemOcc integrates seamlessly with existing 2D-to-3D pipelines without altering backbones or training. Experiments on indoor benchmarks demonstrate MemOcc reduces error propagation by 25% and improves mapping speed over state-of-the-art methods, achieving robust, real-time performance. By selectively retaining reliable evidence and enabling efficient retrieval, MemOcc paves the way for scalable indoor perception in robotics and augmented reality.**

*This research is supported by Artificial Intelligence-National Science and Technology Major Project (2023ZD0121200), the National Natural Science Foundation of China (62441617, 62437001, 62436001), Beijing Natural Science Foundation (L252146), the Key Research Development Program of Jiangsu Province under Grant BE2023016-3, and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDB1350103.

[1]Beihang University, [2]Institute of Automation of the Chinese Academy Sciences, [3]Horizon Robotics, [4]Beijing Jiaotong University, [5]University of Chinese Academy of Sciences. *Corresponding author. (yirongyang@buaa.edu.cn, longteng.guo@ia.ac.cn, wangqb6@outlook.com)

I. INTRODUCTION

Embodied agents equipped with an RGB camera in indoor environments face the fundamental perceptual need to construct holistic 3D scene understanding from egocentric observations. Prior research has primarily focused on multi-view, scene-level embodied perception that parses scenes into semantic or object-level entities [1], [2], [3], [4]. The most common formulations are 3D object bounding-box prediction and semantic occupancy estimation. Based on classic image-based 3D perception models ImVoxelNet [5], increasingly precise and computationally efficient architectures have delivered steady, measurable gains in 3D scene understanding [6], [7], [8].

EmbodiedScan [9] introduces a first-person 3D perception dataset for continuous, scene-level understanding from egocentric streams. Fixed-window pipelines repeatedly project 2D features from the previous t frames into voxels and feed the stacked volume to a 3D decoder. On the one hand, in continuous multi-step observation, scenes often cycle through occlusion and reappearance; repeatedly reprojecting historical features propagates stale errors and lacks a mechanism for selectively retaining reliable voxel evidence, thereby allowing low-confidence or occluded regions to contaminate the global

representation. On the other hand, indoor perception naturally exhibits long-range temporal dependencies due to occlusions and reappearing surfaces. Recomputing each frame from scratch is slow and unstable, whereas naively stacking more frames is inefficient and error-prone. We therefore seek a memory model that accumulates visible evidence from the past and enables fast retrieval when regions are observed again.

In this paper we present **MemOcc**, a memory-augmented 3D occupancy perception framework built around read–write–retrieve operations as shown in Fig. 1. Facing occlusion reappearance cycles that propagate stale errors, we design the Short-Term Memory Cache module (**SmeC**), which performs confidence-aware read–write aggregation to retain reliable voxel evidence and discard transient noise. SmeC module maintains a temporal memory state and a per-voxel confidence map that jointly govern Read-Write Aggregation. SmeC performs read–write aggregation on the previous frames memory, which simultaneously reduces the drawbacks of fixed-window redundant reprojection and inadequate suppression of invisible voxels.

To mitigate the lack of prior reuse and limited generalization over long sequences, we introduce Long-Term Memory Bank (**LmeB**). LmeB compresses the short-term steady state into scene tokens indexed by scene keys, forming a fixed-capacity key–value memory bank for cross-scene reuse. When entering a new scene, similarity search over scene keys retrieves candidate tokens, which are decoded into a voxel prior and conservatively interpolated with the first-frame observation to produce a retrieval-guided initialization. In parallel, a one-off initialization alignment loss aligns the first-frame read-write behavior. To summarize, our contributions include:

(1) We present MemOcc, a drop-in memory layer placed between the 2D encoder and the 3D decoder that preserves tensor shapes and needs no changes to backbone, head, or annotations.

(2) We design Short-Term Memory Cache: a visibility-gated, selective read–write scheme with a per-voxel confidence map that reduces redundant reprojection and stabilizes temporal features.

(3) We introduce Long-Term Memory Bank: scene-token memory with key-based retrieval for cross-scene reuse and retrieval-guided initialization, plus a one-off alignment loss for faster warm-up and better long-horizon generalization.

(4) Extensive experiments on continuous, scene-level indoor 3D occupancy benchmarks show that MemOcc achieves state-of-the-art performance.

II. RELATED WORK

A. Continuous Occupancy Perception

Robotic perception requires the understanding of both 3D geometry and semantics [10]. Most existing approaches mainly focus on 3D bounding box estimation [11], [12], [13], often ignoring finer geometric details and struggling with unfamiliar, out-of-vocabulary objects. To address these challenges, 3D occupancy prediction has emerged as a new task [14], [15], [16], aiming to provide detailed estimates of the occupancy states and semantics within a scene.

Occupancy-based representation models the occupancy state of each voxel in 3D space, offering a dense and structured description of the scene. Unlike traditional point-based or voxel-classification approaches, occupancy prediction focuses on reconstructing the complete 3D occupancy distribution, which can benefit tasks such as object detection, scene completion, and novel view synthesis. Methods like Occ3D [10] and OccNet [17] adopt implicit or explicit voxel occupancy modeling to achieve fine-grained 3D structural understanding. Recently, image-based occupancy approaches such as MVSNeRF [18] and MonoOcc [19] have emerged, aiming to estimate occupancy probabilities directly from multi-view or monocular images. In robot navigation tasks, multi-view 3D occupancy prediction can predict the 3D semantic occupancy of the surrounding environment using image information from multiple views, providing crucial data for navigation and path planning. TPVFormer [20] is a pioneer in this field, utilizing sparse LiDAR labels for supervision and introducing a three-view representation (TPV), which combines BEV and two additional vertical planes. This approach provides basic 3D environment modeling for robot navigation.

B. Long Sequence Modeling in 3D Perception

Long sequence modeling plays a critical role in various perception tasks, especially when handling time series data such as continuous 3D point cloud sequences. ViViT [21] allows Transformers to model temporal and spatial information together, making them well-suited for tasks like 3D object detection from video streams. Chen et al. [22] proposed a method based on autoregressive Transformers. Bai et al. [23] proposed an RSBEV-Mamba framework based on an autoregressive Transformer. This method projects multi-view image features into 3D BEV space and introduces a 3D VMamba module for efficient spatial modeling. Gamba [24] proposes an efficient single-view 3D reconstruction model that processes 3D Gaussian sequences generated from a single image using long sequence modeling techniques. Building upon this, MVGamba [25] introduces a unified 3D generation framework that enhances the quality of 3D content generation through long sequence modeling,

efficiently processing multi-view images using long sequence modeling techniques.

C. Memory-Augmented in 3D Perception

In 3D multi-view perception tasks, external memory can substantially enhance model performance [26]. By leveraging long-term memory, the model can retrieve historical information most relevant to the current observation, thereby improving present perception decisions.

For example, long-term memory helps the model maintain consistency across multi-view data and quickly adapt when the scene changes. Xu et al. [26] proposed a pluggable memory adapter that inserts "write–retrieve–aggregate" memory units into both the point-cloud and image branches, jointly leveraging intra- and cross-modal information. This design endows offline 3D perception models with online temporal enhancement and improves their robustness and accuracy. MAD [27] introduces a sensor-agnostic, plug-and-play memory augmentation module that endows any off-the-shelf 3D detector with long-horizon temporal fusion. Liu et al. [28] propose inserting lightweight, pluggable memory units after the point-cloud and image branches. Following a causal "write–retrieve–aggregate" pipeline, these units maintain and query historical representations.

III. METHOD

A. Continuous Semantic Occupancy Perception

Building on the EmbodiedScan [9] benchmark for continuous semantic occupancy prediction, our goal is to estimate voxel-level occupancy and semantics for indoor scenes from multi-view RGB sequences, emphasizing continuous scene-level representations in enclosed environments. However, Wang et al. [9] adopt a fixed temporal window in which multi-frame 2D features are projected into voxels and then jointly fed to a 3D decoder. While simple, this paradigm has notable limitations: historical frames are repeatedly projected and aggregated, incurring redundant computation; it lacks visibility and uncertainty-aware selective memory, allowing low-confidence or occluded regions to be erroneously written into the historical representation and contaminate it. To this end, we propose MemOcc as shown in Fig. 2, which comprises three components: Short-Term Memory Cache (**SmeC**), Long-Term Memory Bank (**LmeB**), and Cross-Scene Reuse via Long-Short Memory.

B. Short-Term Memory Cache

In order to address stale-error propagation under occlusion–reappearance cycles and the lack of selective retention in fixed-window reprojection, we introduce the SmeC. SmeC aims to maintain a recurrent voxel memory state with a per-voxel confidence map and temporal memory to fuse the current observation into a temporally stable short-term steady state.

Given the extracted image features X_i at time step t as the input of SmeC, the external observation is obtained by 3D spatial mapping, which includes the voxel features $\mathbf{x}_t \in \mathbb{R}^{B\times C\times N_x\times N_y\times N_z}$ and a binary visibility mask $\mathbf{v}_t \in \{0,1\}^{B\times 1\times N_x\times N_y\times N_z}$. Here, N_x, N_y, and N_z denote the grid resolution along the x, y, and z axes, respectively. The SmeC carries voxel memory state from history, comprising a temporal memory state $\mathbf{h}_{t-1} \in \mathbb{R}^{B\times C\times N_x\times N_y\times N_z}$ and a per-voxel confidence map $\mathbf{c}_{t-1} \in \mathbb{R}^{B\times 1\times N_x\times N_y\times N_z}$. The SmeC module fuses the external observation and memory state through read–write aggregation, outputting the fused voxel features $\mathbf{y}_t$ and updated memory states for the t step $\mathbf{x}_t, \mathbf{c}_t$.

Memory Read-Write Aggregation. To quantify how reliably each voxel has been observed in the history ρ_{t-1}, we first apply a projection normalization $\mathcal{P}(\cdot)$ to the previous confidence map $\mathbf{c}_{t-1}$ and then feed it, together with the input context $\mathbf{z}_t = \{\mathbf{x}_t, \mathbf{v}_t, \mathbf{h}_{t-1}, \rho_{t-1}\}$, into the temporal aggregation networks $G_{\mathcal{W}}$ and $G_{\mathcal{R}}$.

$$\boldsymbol{\rho}_{t-1} \;=\; \mathcal{P}(\mathbf{c}_{t-1}/\tau) \in [0,1] \tag{1}$$

$$\boldsymbol{\alpha}_t = \sigma\big(G_{\mathcal{W}}(\mathbf{z}_t)\big) \odot \mathbf{v}_t, \boldsymbol{\beta}_t = \sigma\big(G_{\mathcal{R}}(\mathbf{z}_t)\big) \odot \mathbf{v}_t \tag{2}$$

where τ denotes the coverage saturation threshold, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes Hadamard multiplication, $\boldsymbol{\alpha}_t$ controls the overwrite ratio of the current observation into the history, and $\boldsymbol{\beta}_t$ controls the fusion of history and current when producing the output. The temporal memory state update and the output are:

$$\mathbf{h}_t = \mathbf{v} \odot \big(\alpha_t \odot \mathbf{x}_t + (1-\alpha_t) \odot \mathbf{h}_{t-1}\big) + (1-\mathbf{v}) \odot \mathbf{h}_{t-1} \tag{3}$$

$$\mathbf{y}_t = (1-\beta_t) \odot \mathbf{x}_t + \beta_t \odot \mathbf{h}_{t-1} \tag{4}$$

$$\mathbf{c}_t = \mathcal{P}(\mathbf{c}_{t-1} + \mathbf{v}) \in [\,0,\, c_{\max}] \tag{5}$$

The Read-Write Aggregation makes α_t favor historical accumulation at high-confidence visible voxels and rely more on current observations elsewhere, while β_t smooths transient noise and occlusion-induced flicker at the output. The temporal aggregation networks $G_{\mathcal{W}}$ and $G_{\mathcal{R}}$ is a compact 3D subnetwork composed of $1\times1\times1$ and $3\times3\times3$ convolutions to balance capacity and efficiency. The coverage saturation threshold τ and cap $c_{\max}$ control the growth and saturation of confidence, enabling strong smoothing when priori is abundant and responsiveness when priori is sparse.

With the above design, the SmeC performs read-write weighted voxel aggregation that unifies content-aware fusion weight estimation, geometric constraints from the visibility mask, and uncertainty awareness from per-voxel confidences, thereby stabilizing 3D occupancy features under occlusions and viewpoint changes.

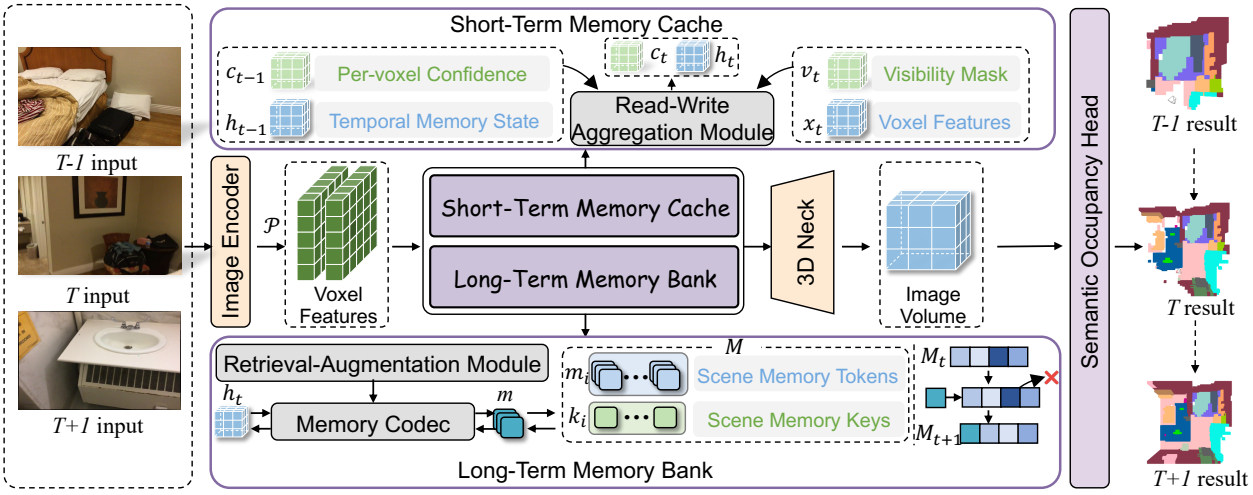

Fig. 2: **Framework diagram of the proposed MemOcc.** MemOcc comprises a Short-Term Memory Cache and a Long-Term Memory Bank.

C. Long-Term Memory Bank

We propose the LmeB with the goal of cross-scene reuse and prior retrieval. The LmeB stores compressed scene tokens indexed by scene keys, which are derived from the steady state of short-term memory. This enables the retrieval of relevant scene priors for new scenes, facilitating faster adaptation by reusing historical context.

The input to the LmeB at time step t comes from the output of the short-term memory $\mathbf{h}_t$ and $\mathbf{c}_t$ at time step t. The Long-Term Memory Bank is updated by compressing the steady memory states, ultimately generating memory tokens composed of scene keys and scene values. Firstly, we construct a scene key–value memory bank $M = \{(k_i, m_i)\}_{i=1}^{M}$, where the key vector $k_i \in \mathbb{R}^{D_k}$ is used for similarity retrieval, and the value vector $m_i \in \mathbb{R}^{D_m}$ is a compressed token of the short-term steady state voxels.

Memory Retrieval-Augmentation. When entering the first frame or view of a new scene, given a unit-normalized query key $k \in \mathbb{R}^{D_k}$, retrieval in the memory bank is performed using cosine similarity:

$$s_i = \mathrm{sim}(k, k_i) = k^\top k_i, \quad i^\star = \arg\max_i s_i. \tag{6}$$

If the hit score satisfies the threshold condition $s_{i^\star} \geq \delta$, then return the value $m_{i^\star}$ of the most similar entry and the similarity $s_{i^\star}$; otherwise, treat as a miss and return the empty set.

Memory Write-Back. After the scene is processed, take the scene keys $\{k_v\}_{v=1}^{V}$ over multiple views or time steps V, average them and re-normalize to obtain the representative key $\bar{k}$,

$$\bar{k} = \mathrm{Norm}\left(\frac{1}{V}\sum_{v=1}^{V} k_v\right) \in \mathbb{R}^{D_k}, \tag{7}$$

and compress the final short-term steady state $h \in \mathbb{R}^{1\times C\times N_x\times N_y\times N_z}$ via the encoder $F_E(\cdot)$ to form a persistent token $m_{\mathcal{L}}$ as shown in the following formula,

$$m_{\mathcal{L}} = F_E(h) \in \mathbb{R}^{D_m} \tag{8}$$

If the current memory size $M < M_{\max}$, append a new item $M_{\mathcal{L}} = \{(\bar{k}_{\mathcal{L}}, m_{\mathcal{L}})\}$ at the tail; when $M = M_{\max}$, select the oldest entry j according to the least recently and overwrite it in place replace $M_j = \{(k_j, m_j)\}$ with $M_{\mathcal{L}} = \{(\bar{k}_{\mathcal{L}}, m_{\mathcal{L}})\}$. To maintain numerical stability and reusability, keys are unit-normalized again before write-back, and values are stored with their original real precision.

By cyclically aggregating scene keys $\bar{k}$, writing back compressed states as memory tokens $m_{\mathcal{L}}$, performing thresholded retrieval to obtain the top token and refreshing entries, the LmeB continuously accrues and updates transferable scene priors within a fixed capacity, supplying downstream modules with reusable tokens that can be decoded for use on demand.

D. Cross-Scene Reuse via Long-Short Memory

At the entry of a new scene, the token $\mathbf{m}_i$ is retrieved from the LmeB into a voxel prior $\mathbf{h}_0$ via decoder $F_{\mathcal{D}}(\cdot)$,

$$\mathbf{h}_0 = F_{\mathcal{D}}(\mathbf{m}_i) \in \mathbb{R}^{D_m}, \tag{9}$$

which is then conservatively fused with the first-frame observation $\mathbf{x}_0 \in \mathbb{R}^{1\times C\times N_x\times N_y\times N_z}$ to form the initialization of the short-term state $\mathbf{h}_{\text{init}}$,

$$\mathbf{h}_{\text{init}} = (1-\tau)\,\mathbf{x}_0 + \tau\,\mathbf{h}_0 \tag{10}$$

This initialization is copied into the hidden state without gradients, and per-voxel confidences are reset, anchoring the recurrence to a cross-scene prior.

To further align the first read-write aggregation with the retrieved prior, we apply an initialization alignment loss at the first frame. We apply an auxiliary loss $\mathcal{L}_{\text{init}}$ to the fused output of the short-term memory $\mathbf{y}_0$ weighted in the total objective.

$$\mathcal{L}_{\text{init}} \;=\; \big\|\mathbf{y}_0 \;-\; \text{sg}\big(\mathbf{h}_{\text{init}}\big)\big\|_1 \tag{11}$$

$$\mathcal{L}_{\text{total}} \;=\; \lambda\,\mathcal{L}_{\text{occ}} + \delta\,\mathcal{L}_{\text{init}} \tag{12}$$

The $\mathcal{L}_{\text{init}}$ softly steers the initial update toward the retrieved prior, suppressing early oscillations under sparse visibility or heavy occlusion, while keeping the retrieval and memory bank outside the gradient path.

IV. Experiment

All experiments were carried out with the PyTorch framework [29] on a cluster of 6 NVIDIA A6000 GPUs, employing mixed-precision computation to improve training efficiency. For MemOcc training, a ResNet50 backbone [30] was used to encode each input view into latent feature representations X_i. Following EmbodiedScan[9] setting, we default to a multi-view sequence length of 10 during training and 20 during testing in the experiments below.

A. Datasets and Metric

EmbodiedScan [9] is a dataset constructed by fusing RGB-D images and corresponding camera poses from existing high-quality 3D indoor datasets such as ScanNet [31], 3RScan [32], and Matterport3D [33]. Specifically, the dataset labels each voxel with a semantic category, generating an occupancy grid map of size 40×40×16, where each occupied voxel belongs to one of 80 predefined semantic categories. The dataset includes 3,113 scene scans for training and 817 scene scans for testing. We use images with a resolution of 480×480 as input and employ the mIoU as the performance metric to measure the overlap between the predicted occupancy regions and the ground truth occupancy regions.

B. Main Results

Table I reports per-class IoU and overall mIoU on the RGB continuous semantic occupancy benchmark, comparing ImVoxelNet, EmbodiedScan, and the proposed MemOcc across four settings: the full EmbodiedScan dataset and its three subsets (ScanNet, Matterport3D, 3RScan). Across all four datasets, MemOcc consistently outperforms the baselines. On the EmbodiedScan benchmark, mIoU increased by 8.60% over ImVoxelNet and by 3.90% over EmbodiedScan; per-class improvements are large, for example toilet increased by 29.09% and floor by 24.95%. On the ScanNet subset, mIoU increased by 9.76% over ImVoxelNet and by 1.84% over EmbodiedScan, with floor increased by 16.08% and bed by 7.50%. On Matterport3D, mIoU increased by 3.67% over ImVoxelNet and by 0.79% over EmbodiedScan, with toilet increased by 7.89% and chair by 6.49%. On 3RScan, mIoU increased by 6.62% over ImVoxelNet and by 1.07% over EmbodiedScan, with bed increased by 4.85% and chair by 4.54%. Overall, MemOcc delivers consistent gains on both overall metrics and representative categories across domains, highlighting its strong robustness and generalization.

As shown in Table II, to investigate the impact of different training sequence lengths on the model, we conduct experiments separately on the ScanNet and Matterport3D subsets. The result shows that On ScanNet with Seq 10, mIoU increased by 9.76% over ImVoxelNet and by 1.84% over EmbodiedScan; with Seq 20, mIoU increased by 7.78% over ImVoxelNet and by 0.95% over EmbodiedScan. Increasing the sequence length from 10 to 20 on ScanNet decreased MemOcc by 1.83% and decreased EmbodiedScan by 0.94% while ImVoxelNet increased by 0.15%. On MatterPort3D with Seq 10, mIoU increased by 3.67% over ImVoxelNet and by 0.79% over EmbodiedScan; with Seq 20, mIoU increased by 4.24% over ImVoxelNet and by 0.07% over EmbodiedScan. Moving from Seq 10 to Seq 20 on MatterPort3D increased MemOcc by 0.47% and increased EmbodiedScan by 1.19% while ImVoxelNet decreased by 0.10%. Overall, MemOcc remains the best performer across both sequence lengths and datasets, and the sequence-length effect is domain dependent, with shorter sequences favored on ScanNet and longer sequences slightly preferred on MatterPort3D.

C. Ablation Studies

Overall Ablation. Table III shows that the main gain comes from the Short-Term Memory Cache: mIoU increased from 6.66 to 16.07, a rise of 9.41%. Adding the Long-Term Memory Bank on top of SmeC produced a further increase to 16.17, an additional 0.10%. The full model MemOcc achieved 16.42, increasing by 0.25% over SmeC+LmeB, by 0.35% over SmeC, and by 9.76% over the baseline, indicating that short-term memory delivers the dominant improvement while long-term memory provides a smaller but consistent boost.

Ablation of SmeC. We assess the contributions of three key components in the SmeC module as shown in Table IV. Among them, *Write-only* denotes writing the current observation into the history without fusing it with the current input, while *Fixed* $\boldsymbol{\beta}_t$=0.3 indicates that the fusion probability between the historical record and the current observation is fixed at 0.3. The full SmeC reaches 16.07 mIoU. Removing the temporal memory state $\mathbf{h}_t$ reduces performance to 6.05, decreased by 10.02%, because there is no persistent state to accumulate reliable voxel evidence across frames, which makes predictions vulnerable to occlusion–reappearance cycles and reprojection noise. Removing the per-voxel

TABLE I: Results of MemOcc on the RGB continuous semantic occupancy benchmark. We train and evaluate on the EmbodiedScan dataset [9] and its three subsets (Scannet[31], MatterPort3D[33], 3RScan[32]). The multi-view image sequence length is set to 10 during training and 20 during testing.

Method	floor	toilet	bed	wall	curtain	chair	couch	table	shelf	plant	door	cabinet	empty	mIoU
ImVoxelNet [5]	48.38	21.32	14.09	14.65	11.61	12.20	10.53	14.85	11.22	6.03	9.20	6.53	34.95	5.73
EmbodiedScan [9]	34.10	29.78	24.46	30.24	19.80	26.46	35.19	41.60	16.22	17.26	25.53	9.49	39.09	10.43
MemOcc	**59.05**	**58.87**	**36.46**	**33.67**	**32.91**	**31.91**	30.00	28.89	**24.76**	**21.94**	20.68	**19.64**	**52.24**	**14.33**
ImVoxelNet [5]	68.15	14.65	11.64	13.82	10.01	13.90	11.62	12.10	11.99	5.37	7.83	5.17	36.62	6.66
EmbodiedScan [9]	54.35	55.84	33.84	34.01	35.23	31.48	28.55	29.96	25.60	20.97	21.87	22.59	51.67	14.58
MemOcc	**70.43**	**62.03**	**41.34**	31.89	34.93	**38.20**	**29.26**	**32.42**	**31.68**	7.20	18.53	19.17	**54.31**	**16.42**
ImVoxelNet [5]	38.74	21.32	14.09	14.65	11.61	2.95	6.68	3.53	2.24	1.86	8.64	2.59	25.36	3.27
EmbodiedScan [9]	44.82	22.07	25.82	24.40	5.04	7.72	20.88	9.39	1.14	6.05	11.19	6.96	40.58	6.15
MemOcc	42.98	**29.96**	**29.51**	**24.63**	5.74	**14.21**	**23.67**	**11.68**	**7.62**	**6.13**	**16.81**	**12.44**	39.8	**6.94**
ImVoxelNet [5]	58.96	2.04	3.54	9.70	6.79	5.48	3.68	4.56	3.86	6.79	1.55	4.08	34.55	3.15
EmbodiedScan [9]	57.54	27.43	12.27	29.73	31.57	22.56	20.39	14.57	16.33	10.66	2.30	19.82	55.02	8.70
MemOcc	**59.83**	16.86	**17.12**	**29.99**	**35.71**	**27.10**	**22.82**	14.13	15.13	9.62	**4.86**	17.01	**55.42**	**9.77**

TABLE II: Impact of training sequence length on semantic occupancy performance. The table compares the mIoU of MemOcc and the baselines on the ScanNet and Matterport3D subsets under different training lengths.

Methods	ScanNet		MatterPort3D	
	$Seq = 10$	$Seq = 20$	$Seq = 10$	$Seq = 20$
ImvoxelNet[5]	6.66	6.81	3.27	3.17
EmbodiedScan[9]	14.58	13.64	6.15	7.34
MemOcc	**16.42**	**14.59**	**6.94**	**7.41**

TABLE III: Ablation of MemOcc components on the Scannet val Dataset. The table compares the baseline, SmeC, SmeC with LmeB, and the full MemOcc.

Methods	mIoU
baseline	6.66
+SmeC	16.07
+SmeC + LmeB	16.17
MemOcc	**16.42**

TABLE IV: Ablation of the key components in the SmeC module on the Scannet dataset.

Variant	mIoU
SmeC	**16.07**
w/o Temporal memory state	6.05
w/o Per-voxel confidence	6.40
Write-only	14.84
Fixed $\boldsymbol{\beta}_t$=0.3	6.48

TABLE V: Ablation of key hyperparameters of the Long-Term Memory Bank on the ScanNet dataset.

Capacity	Token dim	Similarity	mIoU
50	1024	0.70	15.98
100	2048	0.72	15.43
300	2048	0.72	**16.17**
600	4096	0.75	15.52

confidence map $\mathbf{c}_t$ reduces performance to 6.40, decreased by 9.67%, since updates cannot down-weight low-confidence or invisible voxels and noisy observations overwrite the steady state. Switching to write-only lowers mIoU to 14.84, decreased by 1.23%, because the system cannot read and aggregate the stable state before writing, so transient inputs overwrite useful history and induce drift. Using a fixed 0.3 rule yields 6.48, decreased by 9.59%, as a static update ratio cannot adapt to spatial and temporal variability in visibility and confidence, leading to under-updates in informative regions or overwrites with unreliable inputs. These results indicate that both $\mathbf{h}_t$ and $\mathbf{c}_t$ are essential and that the read–write mechanism enables robust, content-aware updates.

Ablation of LmeB. Table V examines the influence of long-term memory capacity, token dimensionality, and similarity threshold on performance. Moving from capacity 50 with dim 1024 and similarity 0.70 to capacity 100 with dim 2048 and similarity 0.72 decreased mIoU by 0.55%, indicating that enlarging the bank without stronger selection introduces redundancy and retrieval confusion while a higher threshold reduces recall of useful priors. Increasing capacity to 300 while keeping dim 2048 and similarity 0.72 increased mIoU by 0.74% over the 100 configuration and by 0.19% over the 50 configuration, suggesting a balance where coverage and precision are both adequate. Further pushing to capacity 600 with dim 4096 and similarity 0.75 decreased mIoU by 0.65% relative to 300 and by 0.46% relative to 50, likely because a larger bank and higher-dimensional

TABLE VI: Ablation of the retrieval similarity threshold and the initialization alignment weight for the long-term memory bank on the ScanNet dataset.

τ_{ret}	δ_{init}	floor	toilet	bed	empty	mIoU
0.0	0.00	72.11	58.42	40.87	55.37	16.01
0.4	0.00	69.69	59.96	40.90	54.44	15.60
0.5	0.01	70.43	62.03	41.34	54.31	**16.42**
0.6	0.01	70.69	57,65	40.32	54.06	15.43
0.8	0.01	69.56	59.86	41.59	53.98	14.98

tokens amplify matching noise and false positives, and a stricter threshold suppresses retrieval coverage needed for robust initialization.

Table VI evaluates sensitivity to the retrieval similarity threshold τ_{ret} and the initialization alignment weight δ_{init}. Without either component (τ_{ret}=0.0, δ_{init}=0), the baseline achieves $\text{mIoU} = 16.01$. Raising only the retrieval weight to 0.4% while keeping δ_{init}=0 reduces mIoU by 0.41%, suggesting that emphasizing the historical token alone can inject mismatch noise. Setting τ_{ret}=0.5 and adding a mild alignment loss (δ_{init}=0.01) lifts mIoU by 0.41%, the best result, indicating that *moderate retrieval plus light alignment* stabilizes warm-start and yields overall gains. Further increases in τ_{ret} to 0.6% and 0.8% (with δ_{init}=0.01) lower mIoU by 0.99% and 1.44%, implying that overly strong history injection dilutes current observations and weakens generalization.

TABLE VII: Ablation of cross-scene reuse via long–short memory on the Scannet [31] dataset.

Setting	floor	toilet	bed	empty	mIoU (t = 0 - 3)
w/o-TGI	68.86	19.64	12.05	36.42	6.53
TGI	69.33	20.88	12.11	36.76	**6.59**
Neg-TGI	71.82	20.47	12.33	36.51	6.27

Ablation of Cross-Scene Reuse via Long-Short Memory. Token-guided Initialization (TGI) is applied at the first frame of a sequence to decode an initial voxel memory from a long-term token and to fuse it with the current observation, which serves as the initial state of the short-term memory unit and enables cross-scene reuse. To investigate the robustness of Token-guided Initialization, we design three groups of experiments: without using this prior (w/o TGI), initializing the first frame with a token from the same scene, and initializing the first frame with a Neg-TGI token from another scene with low similarity. As shown in Table VII, we evaluate the first 4 frames (t=0-3) of each sequence to assess the model's immediate performance upon entering a scene. The result shows that Using TGI increased mIoU by 0.06% over w/o-TGI, with toilet increased by 1.24%, floor increased by 0.47%, bed increased by 0.06%, and empty increased by 0.34%. The small but consistent gains indicate that decoding a prior token and interpolating it with the first observation provides a better warm start while the short horizon t=0–3 and conservative interpolation limit the aggregate lift.

Using a low-similarity token in Neg-TGI decreased mIoU by 0.32% relative to TGI and by 0.26% relative to w/o-TGI. The decline arises because a mismatched prior biases the initial memory toward incorrect geometry and semantics, which contaminates early writes before the gating can recover. Although floor increased by 2.49% relative to TGI and by 2.96% relative to w/o-TGI, toilet decreased by 0.41% relative to TGI and empty decreased by 0.25% relative to TGI, and additional unlisted categories also degrade, yielding a lower overall mIoU. These results show that TGI is beneficial when retrieval is similarity-aware, whereas incorrect tokens harm early-scene stability.

TABLE VIII: Cross-dataset evaluation. Trained on ScanNet dataset, tested on 3RScan [32] and MatterPort3D [33] datasets without finetune.

Method	floor	toilet	bed	empty	mIoU
ImvoxelNet	16.75	2.04	3.12	0.22	1.05
MemOcc	**35.15**	**17.97**	**25.01**	**27.47**	**2.24**
ImvoxelNet	14.17	6.83	1.36	16.19	0.87
MemOcc	**24.12**	**28.56**	**32.40**	**19.85**	**3.47**

D. Cross-Scene Transfer

Embodied agents performing multi-view occupancy perception are typically trained on a single dataset and then deployed in new environments with markedly different styles. Table VIII evaluates cross-dataset generalization when training on ScanNet and testing without finetuning on 3RScandataset and MatterPort3D datasets. On the first target dataset, mIoU increased by 1.19% over ImVoxelNet, with the largest category gains on empty increased by 27.25% and bed increased by 21.89%. On the second target dataset, mIoU increased by 2.60%. These results indicate that short-term selective read–write stabilizes structural surfaces such as floor and empty under domain shift, while token-guided initialization from the long-term memory facilitates transferring object semantics such as bed and toilet across scenes.

V. CONCLUSIONS

We presented MemOcc, a memory-augmented framework for indoor embodied continuous 3D semantic occupancy. MemOcc couples a Short-Term Memory Cache that performs visibility-gated, confidence-aware read–write to accumulate reliable voxel evidence and suppress occlusion-induced flicker, with a Long-Term Memory Bank that stores key-indexed scene tokens to enable retrieval-guided initialization and cross-scene reuse. Experiments on EmbodiedScan and its subsets demonstrate consistent gains in long-horizon stability and efficiency, achieving state-of-the-art performance.

REFERENCES

[1] H. Zhou, J. Shao, L. Xu, D. Bai, W. Qiu, B. Liu, Y. Wang, A. Geiger, and Y. Liao, "Hugs: Holistic urban 3d scene understanding via gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 336–21 345.

[2] L. Kong, X. Xu, J. Ren, W. Zhang, L. Pan, K. Chen, W. T. Ooi, and Z. Liu, "Multi-modal data-efficient 3d scene understanding for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 5, pp. 3748–3765, 2025.

[3] W. Xu, C. Shi, S. Tu, X. Zhou, D. Liang, and X. Bai, "A unified framework for 3d scene understanding," *Advances in Neural Information Processing Systems*, vol. 37, pp. 59 468–59 490, 2024.

[4] H. Li, D. Zhang, Y. Dai, N. Liu, L. Cheng, J. Li, J. Wang, and J. Han, "Gp-nerf: Generalized perception nerf for context-aware 3d scene understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 21 708–21 718.

[5] D. Rukhovich, A. Vorontsova, and A. Konushin, "Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2397–2406.

[6] Y. Zhang, Y. Wang, Y. Cui, and L.-P. Chau, "3dgeodet: General-purpose geometry-aware image-based 3d object detection," *IEEE Transactions on Multimedia*, 2025.

[7] P. Shi, W. Wu, and A. Yang, "Mpvf: Multi-modal 3d object detection algorithm with pointwise and voxelwise fusion," *Algorithms*, vol. 18, no. 3, p. 172, 2025.

[8] T. Tu, S.-P. Chuang, Y.-L. Liu, C. Sun, K. Zhang, D. Roy, C.-H. Kuo, and M. Sun, "Imgeonet: Image-induced geometry-aware voxel representation for multi-view 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6996–7007.

[9] T. Wang, X. Mao, C. Zhu, R. Xu, R. Lyu, P. Li, X. Chen, W. Zhang, K. Chen, T. Xue *et al.*, "Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 757–19 767.

[10] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 318–64 330, 2023.

[11] Y. Jia, J. Liu, S. Chen, C. Gu, Z. Wang, L. Luo, X. Li, P. Wang, Z. Wang, R. Zhang *et al.*, "Lift3d policy: Lifting 2d foundation models for robust 3d robotic manipulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 347–17 358.

[12] S. Chen, R. Garcia, I. Laptev, and C. Schmid, "Sugar: Pre-training 3d visual representations for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 049–18 060.

[13] F. Ma, X. Yan, G. Zhao, X. Xu, Y. Liu, J. Ma, and M. Liu, "Every dataset counts: Scaling up monocular 3d object detection with joint datasets training," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 11 574–11 580.

[14] Y. Huang, W. Zheng, B. Zhang, J. Zhou, and J. Lu, "Selfocc: Self-supervised vision-based 3d occupancy prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 19 946–19 956.

[15] H. Liu, Y. Chen, H. Wang, Z. Yang, T. Li, J. Zeng, L. Chen, H. Li, and L. Wang, "Fully sparse 3d occupancy prediction," in *European Conference on Computer Vision*. Springer, 2024, pp. 54–71.

[16] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 729–21 740.

[17] A. Ahmed, Z. Xiaoyang, M. H. Tunio, M. H. Butt, S. A. Shah, Y. Chengxiao, F. A. Pirzado, and A. Aziz, "Occnet: Improving imbalanced multi-centred ovarian cancer subtype classification in whole slide images," in *2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 2023, pp. 1–8.

[18] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su, "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 124–14 133.

[19] Y. Zheng, X. Li, P. Li, Y. Zheng, B. Jin, C. Zhong, X. Long, H. Zhao, and Q. Zhang, "Monoocc: Digging into monocular semantic occupancy prediction," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 18 398–18 405.

[20] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9223–9232.

[21] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[22] C. Ziwen, H. Tan, K. Zhang, S. Bi, F. Luan, Y. Hong, L. Fuxin, and Z. Xu, "Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 4349–4359.

[23] B. Lin, Z. Zou, and Z. Shi, "Rsbev-mamba: 3d bev sequence modeling for multi-view remote sensing scene segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[24] Q. Shen, Z. Wu, X. Yi, P. Zhou, H. Zhang, S. Yan, and X. Wang, "Gamba: Marry gaussian splatting with mamba for single-view 3d reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[25] X. Yi, Z. Wu, Q. Shen, Q. Xu, P. Zhou, J.-H. Lim, S. Yan, X. Wang, and H. Zhang, "Mvgamba: Unify 3d content generation as state space sequence modeling," *Advances in Neural Information Processing Systems*, vol. 37, pp. 7580–7607, 2024.

[26] X. Xu, C. Xia, Z. Wang, L. Zhao, Y. Duan, J. Zhou, and J. Lu, "Memory-based adapters for online 3d scene perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 604–21 613.

[27] B. Agro, S. Casas, P. Wang, T. Gilles, and R. Urtasun, "Mad: Memory-augmented detection of 3d objects," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1449–1460.

[28] H. Liu, B. Laeng, and N. O. Czajkowski, "Visual short-term memory is modulated by 3d depth in stereopsis," *Attention, Perception, & Psychophysics*, vol. 87, no. 8, pp. 2265–2274, 2025.

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

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

[33] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.