Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

FreeOcc: Training-Free Embodied Open-Vocabulary Occupancy Prediction

Zeyu Jiang*[1] Changqing Zhou*[1] Xingxing Zuo[2] Changhao Chen[1]✉
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Mohamed bin Zayed University of Artificial Intelligence

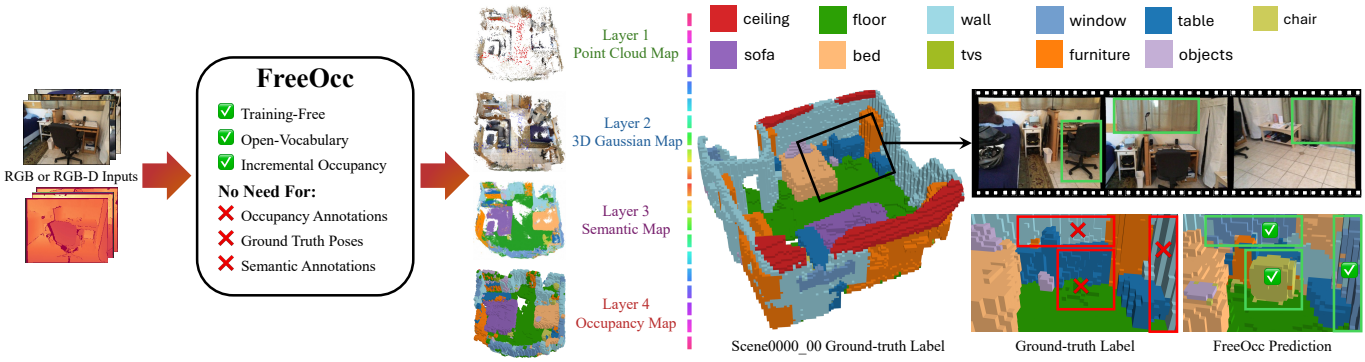

Fig. 1: FreeOcc is a training-free paradigm for open-vocabulary occupancy prediction. It eliminates the need for occupancy, pose, and semantic annotations, and incrementally constructs four-layer maps using only monocular or RGB-D image sequences. The right panel illustrates the benefit of open-vocabulary reasoning on EmbodiedOcc-ScanNet: the green boxes corresponding to "window" and "chair" are correctly identified and localized by FreeOcc, whereas the EmbodiedOcc-ScanNet ground-truth occupancy labels (red boxes) coarsely classify them as "wall" and "floor" respectively, despite clear visual evidence.

***Abstract*—Existing learning-based occupancy prediction methods rely on large-scale 3D annotations and generalize poorly across environments. We present FreeOcc, a training-free framework for open-vocabulary occupancy prediction from monocular or RGB-D sequences. Unlike prior approaches that require voxel-level supervision and ground-truth camera poses, FreeOcc operates without 3D annotations, pose ground truth, or any learning stage. FreeOcc incrementally builds a globally consistent occupancy map via a four-layer pipeline: a SLAM backbone estimates poses and sparse geometry; a geometrically consistent Gaussian update constructs dense 3D Gaussian maps; open-vocabulary semantics from off-the-shelf vision–language models are associated with Gaussian primitives; and a probabilistic Gaussian-to-occupancy projection produces dense voxel occupancy. Despite being entirely training-free and pose-agnostic, FreeOcc achieves over $2\times$ improvements in IoU and mIoU on EmbodiedOcc-ScanNet compared to prior self-supervised methods. We further introduce ReplicaOcc, a benchmark for indoor open-vocabulary occupancy prediction, and show that FreeOcc transfers zero-shot to novel environments, substantially outperforming both supervised and self-supervised baselines. Project page: https://the-masses.github.io/freeocc-web/.**

I. Introduction

The ability to construct and understand the environment from egocentric observations is fundamental to embodied lifelong autonomy. Beyond geometric reconstruction, robots require scene representations that capture dense structure, semantic completeness, and global spatial consistency to support navigation and interaction in open environments [9]. Point clouds arise naturally in depth sensing, structure-from-motion [65], and SLAM [4, 41], and have long served as a core representation for robotic perception [6]. However, their unstructured nature and irregular sampling limit their effectiveness for downstream reasoning. 3D Gaussian Splatting (3DGS) [29] addresses these limitations by augmenting each 3D sample with anisotropic extent and opacity, providing a compact continuous representation suitable for differentiable rendering. Despite its advantages, 3DGS is typically optimized under photometric supervision, often resulting in inaccurate geometry, depth distortions, and view-inconsistent artifacts [52, 66, 48]. Moreover, geometric boundaries are commonly extracted via heuristic density thresholds, yielding ambiguous or inconsistent object extents [10].

In contrast, occupancy maps discretize space into free, occupied, and unknown regions, providing explicit geometric boundaries and supporting incremental updates [17, 58]. Since these boundaries are directly used for collision checking and motion planning, their accuracy is critical for safety and task performance [11]. Motivated by the complementary strengths of Gaussians and occupancy, recent works combine 3D Gaussian primitives with voxelized occupancy representations [25, 22, 23], leading to the embodied occupancy prediction task [72], which estimates semantic occupancy volumes from Gaussians built from egocentric observations.

Embodied occupancy prediction has advanced rapidly, achieving strong geometric and semantic accuracy in indoor environments and, in some cases, real-time inference [72, 69, 79, 37]. However, most existing methods rely on fully su-

*Equal contribution. ✉ Corresponding author.

pervised voxel-level annotations and assume accurate camera poses at inference [76, 8]. Such supervision is expensive, requiring large-scale reconstruction and labeling, and methods trained in this regime often generalize poorly beyond the training distribution. While recent self-supervised approaches reduce annotation requirements and introduce open-vocabulary semantics via vision–language models [25, 12, 53, 50], learning-based methods still depend on accurate poses during training and inference and tend to overfit to specific scenes, viewpoints, or sensor configurations, leading to significant performance degradation in novel environments.

To address these limitations, we propose FreeOcc, the first training-free framework for open-vocabulary occupancy prediction, which incrementally constructs a globally consistent occupancy map through a four-layer mapping pipeline. **Layer 1:** A SLAM backbone processes monocular or RGB-D image sequences to estimate camera poses and build sparse point cloud maps. **Layer 2:** We construct dense 3D Gaussian maps using SLAM-guided point initialization and a geometrically consistent Gaussian update strategy, ensuring structural fidelity and long-term global consistency. **Layer 3:** Open-vocabulary semantic features extracted from pre-trained vision–language models (VLMs) are incrementally associated with Gaussian primitives, enabling language-based querying without voxel-level supervision. **Layer 4:** A probabilistic Gaussian-to-occupancy projection aggregates geometric and semantic evidence into a discrete voxel grid, yielding a dense occupancy map with open-vocabulary semantics. On the EmbodiedOcc-ScanNet benchmark, FreeOcc achieves over $\mathbf{2}\times$ improvements in both IoU and mIoU compared to self-supervised methods. We further introduce ReplicaOcc, a new benchmark for evaluating generalization in indoor open-vocabulary occupancy prediction, on which FreeOcc demonstrates strong zero-shot generalization, significantly outperforming both supervised and self-supervised learning-based baselines across all metrics.

In summary, we make the following contributions:

- We propose **FreeOcc**, a training-free framework for open-vocabulary occupancy prediction that addresses the poor generalization of existing occupancy prediction methods caused by dataset-specific training and closed-set supervision.
- We identify the geometric ambiguity arising from decoupled 3DGS-SLAM optimization and introduce a novel globally consistent Gaussian update strategy, where geometrically anchored updates produce more spatially consistent 3D scene representations for occupancy mapping.
- We present **ReplicaOcc**, a new benchmark for evaluating generalization in open-vocabulary occupancy prediction. Extensive experiments demonstrate that FreeOcc significantly outperforms prior self-supervised methods and substantially improves generalization compared to learning-based approaches.

II. Related Work

A. Fully Supervised Occupancy Prediction

Vision-based occupancy prediction has been widely studied, initially in outdoor autonomous driving benchmarks [62, 22, 23, 70, 38] and more recently in indoor and embodied environments [5, 75, 59, 76, 72, 69, 81, 82]. Fully supervised methods lift 2D image features into 3D representations using depth distributions, ray-based projection, or volumetric aggregation [5, 75, 76, 39, 51]. Transformer-based volumetric models capture long-range dependencies [56], while sparsity-aware designs prune empty regions and process sparse voxels via sparse convolutions [62] or efficient Transformers [38, 40]. Despite strong performance, these approaches rely on dense voxel-level annotations that are expensive to obtain and difficult to scale. Moreover, full supervision often leads to limited generalization to novel scenes or sensor configurations.

B. Weakly Supervised Occupancy Prediction

Weakly supervised methods reduce annotation costs by learning occupancy from indirect supervision such as 2D segmentation, sparse LiDAR, or pseudo-labels [32]. Several approaches optimize 3D occupancy using 2D-only supervision via differentiable rendering or image-space distillation [47, 2, 36]. Others construct approximate 3D supervision by aggregating sparse LiDAR points [80, 13, 67, 36] or enforce multi-view consistency through image-plane reprojection [21, 78, 26, 1, 12]. While weak supervision alleviates labeling requirements, most methods assume accurate camera poses, operate in fixed domains, and perform offline inference, limiting their applicability to embodied agents that require online, incremental mapping [72].

C. 3D Gaussian Splatting SLAM

3D Gaussian Splatting SLAM (3DGS-SLAM) jointly estimates camera poses and optimizes continuous Gaussian maps, and has recently attracted significant attention [73, 19, 27, 43, 77, 30, 31, 33]. Its continuous, differentiable, and compact representation [29] makes it well-suited as a geometric carrier for occupancy prediction [22]. Extensions to semantic and open-set 3DGS-SLAM further enable semantic self-supervision and open-vocabulary reasoning [35, 24, 34, 74]. However, existing 3DGS-SLAM methods primarily optimize photometric objectives for view synthesis, often resulting in spatial inconsistencies that limit voxel-level completion [27, 15, 19]. In contrast, our work introduces a geometrically consistent 3DGS update strategy and leverages it as a prior for training-free, open-vocabulary occupancy prediction, bridging continuous surface mapping and discrete volumetric reasoning.

III. Problem Statement

Task Overview. FreeOcc addresses the problem of *embodied* semantic occupancy prediction [72, 69, 79]. In contrast to monocular scene completion methods that infer a 3D occupancy map from a single RGB image [59, 68, 76], the embodied setting requires a robot to *incrementally* construct a

TABLE I: Comparison of supervision, inference inputs, and outputs across methods. **O**: human-labeled occupancy; **S**: human-labeled semantics; **P**: human-labeled poses; **D**: depth; **R**: RGB images. Outputs include **OV** (open-vocabulary semantics) and **OCC** (occupancy prediction).

Method	Train	Infer	Output
Fully supervised			
EmbodiedOcc [72]	O,S,P,D,R	P,R	OCC
EmbodiedOcc++ [69]	O,S,P,D,R	P,R	OCC
RoboOcc [79]	O,S,P,D,R	P,R	OCC
Self-supervised			
GaussTR [25]	P,R	P,R	OV,OCC
GaussianOCC [12]	P,R	P,R	OV,OCC
Training-free			
FreeOcc (mono)	–	R	OV,OCC
FreeOcc (rgbd)	–	D,R	OV,OCC

globally consistent semantic occupancy map from egocentric observations while actively exploring the environment.

Formally, given a stream of RGB observations $\mathcal{I}_{1:T} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_T\}$, our goal is to estimate a global 3D semantic occupancy field $\mathbf{O}_T \in \mathbb{R}^{X\times Y\times Z\times C}$ in an online manner. Here, (X, Y, Z) denote the spatial resolution of the scene volume, and C is the number of semantic categories. Each voxel encodes both geometric occupancy (occupied or free) and semantic evidence, representing the environment observed up to time T.

Key Differences from Prior Work. As summarized in Tab. I, existing occupancy prediction approaches can be broadly categorized according to the supervision and prior information required during training and inference:

- *Fully supervised* methods [72, 69, 79] rely on dense voxel-level annotations, which are costly to obtain and typically require large-scale 3D reconstruction followed by manual or semi-automatic labeling pipelines.
- *Self-supervised* methods [25, 12] reduce dependence on voxel annotations but still assume *known camera poses* during both training and inference. Moreover, as learning-based approaches, both supervised and self-supervised methods often suffer from limited cross-scene generalization, leading to degraded zero-shot performance in previously unseen environments, as demonstrated in Sec. V-C.

In contrast, **FreeOcc** is a *training-free* approach that requires neither semantic annotations nor prespecified camera poses and directly predicts embodied occupancy from an incoming video stream. Furthermore, since RGB and depth sequences are commonly available in robotic systems, FreeOcc supports two inference modes: monocular RGB and RGB-D, enabling flexible deployment across diverse sensing platforms.

IV. Methodology

In this section, we first present an overview of the FreeOcc system architecture in Sec. IV-A. We then detail the geometrically consistent 3D Gaussian mapping process in Sec. IV-B. Next, we introduce the open-vocabulary semantic association module in Sec. IV-C, which injects language-aligned semantics into the Gaussian map to form language-embedded (LE) Gaussians. Finally, we describe how the LE-Gaussian representation is converted into a volumetric occupancy map in Sec. IV-D, enabling open-vocabulary querying with arbitrary text prompts.

A. Overall Architecture of FreeOcc

The overall architecture of FreeOcc is illustrated in Fig. 2. At a high level, FreeOcc is an online, training-free system that incrementally constructs an open-vocabulary 3D occupancy map from monocular or RGB-D image streams. Specifically, FreeOcc explicitly maintains multi-level scene representations and continuously refines them as new observations arrive, in a fully streaming fashion. It consists of four tightly coupled components: **Layer 1** a SLAM backbone for globally consistent camera pose estimation and geometric reconstruction, **Layer 2** a geometrically anchored 3D Gaussian mapping module that lifts SLAM point clouds into a continuous scene representation, **Layer 3** an open-vocabulary semantic association module that assigns language-aligned embeddings to Gaussian primitives, and **Layer 4** a Gaussian-to-occupancy projection module that converts the Gaussian representation into a volumetric occupancy field. We describe each component in detail below.

Geometric Correspondence using SLAM Backbone. FreeOcc first processes incoming observations using a SLAM backbone to estimate camera poses and sparse 3D geometry. In principle, the proposed framework is compatible with arbitrary SLAM systems. In this work, we adopt DROID-SLAM [64] due to its strong global geometric consistency and robustness under monocular input. Unlike feed-forward, model-based SLAM approaches such as MASt3R-SLAM [46] or VGGT-SLAM [42], DROID-SLAM does not rely on explicit 3D supervision from structure-from-motion (SfM) pipelines [54] during training of its optical flow network [63]. By jointly optimizing over long temporal windows, DROID-SLAM produces globally consistent camera poses $\mathcal{T}_{1:T} = \{\mathbf{T}_1, \ldots, \mathbf{T}_T\}$ and an accumulated set of 3D points $\mathcal{P}_{1:T} = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^{N_T}$ which together provide a stable spatial reference for downstream mapping. This global consistency allows all subsequent modules to operate in a unified coordinate frame and is critical for mitigating geometric drift in long-horizon mapping.

Geometrically Consistent 3D Gaussian Construction. Given globally consistent camera poses $\mathcal{T}_{1:T}$ and 3D points $\mathcal{P}_{1:T}$ from SLAM, FreeOcc incrementally maintains a set of 3D Gaussian primitives $\mathcal{G} = \{G_i\}_{i=1}^{N_T}$.We update the Gaussians while preserving multi-view geometric consistency rather than optimizing for novel-view synthesis as in prior 3DGS-SLAM [16]. The resulting Gaussian map bridges sparse SLAM geometry and dense volumetric occupancy reasoning.

Open-Vocabulary Semantic Association. We extract open-vocabulary semantics from 2D observations using pre-trained vision–language models [53, 57] and fuse them into Gaussian primitives via geometric correspondence. The aggregated semantics are propagated to the occupancy grid for language-driven querying without voxel-level supervision.

Gaussian-to-Occupancy Projection. Finally, the maintained Gaussian set $\mathcal{G}$ is projected into a discrete occupancy field $\mathbf{O} \in$

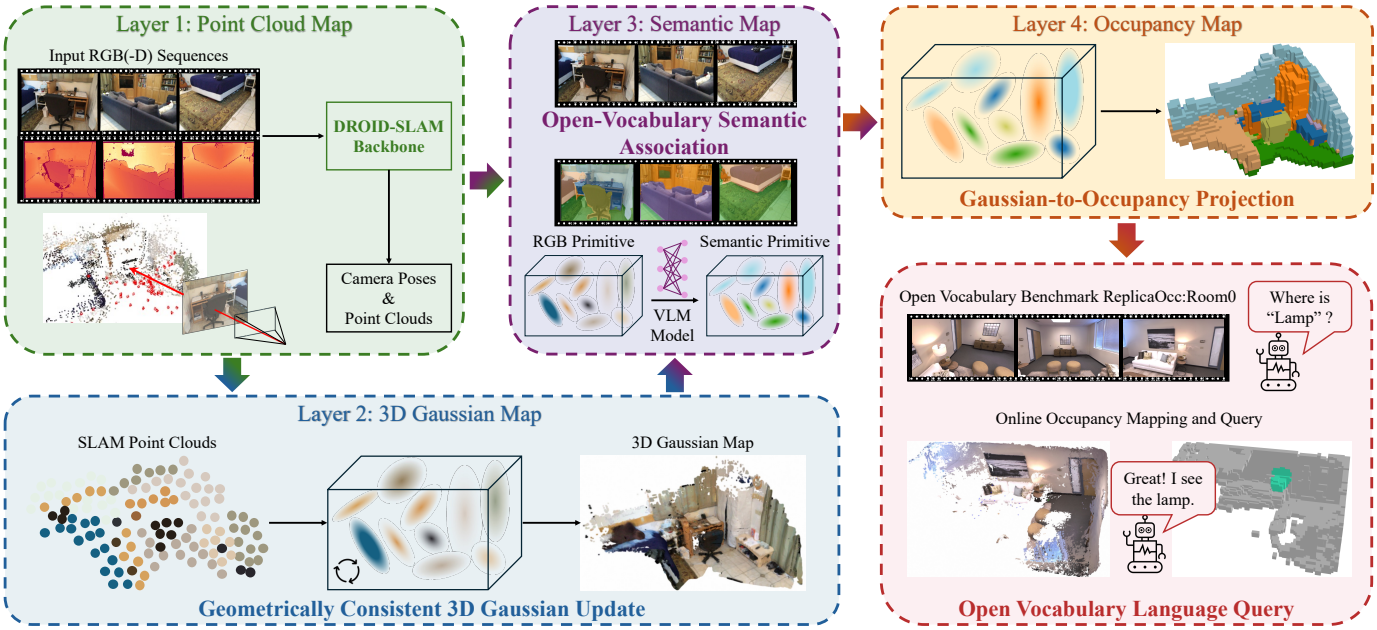

Fig. 2: **Framework Overview of FreeOcc.** FreeOcc incrementally constructs a multi-layer map for online open-vocabulary occupancy prediction. **Layer 1:** A SLAM backbone processes monocular or RGB-D image sequences to estimate camera poses and sparse/semi-dense point cloud maps. **Layer 2:** Dense 3D Gaussian Splatting (3DGS) maps are constructed via SLAM-guided point initialization and a geometrically consistent Gaussian update strategy. **Layer 3:** Open-vocabulary semantic features are associated with Gaussian primitives using a vision–language model, forming a language-embedded 3D Gaussian semantic map. **Layer 4:** The semantic Gaussian map is projected into a dense voxel occupancy representation through probabilistic Gaussian-to-occupancy splatting, enabling online open-vocabulary querying and semantic localization in 3D scenes.

$\mathbb{R}^{X\times Y\times Z\times C}$ by aggregating the probabilistic spatial support of nearby Gaussian primitives.

B. Geometrically Consistent 3D Gaussian Construction

Geometric Ambiguity Problem. Recent 3DGS-SLAM approaches grounded in point-based SLAM systems—such as Photo-SLAM [19] built upon ORB-SLAM [45], and DROID-Splat [64] built upon DROID-SLAM [64]—naturally inherit the localization accuracy and efficiency of classical SLAM pipelines. However, these methods share a fundamental property: the Gaussian scene representation is optimized independently of the point-based SLAM map. In such decoupled systems, Gaussian parameters are primarily updated to maintain rendering consistency, while the SLAM backend enforces geometric consistency [18].

Following standard 3DGS formulations, we represent the scene as a set of language-embedded Gaussian primitives, each parameterized as $G_i = (\boldsymbol{\mu}_i, \mathbf{s}_i, \mathbf{r}_i, o_i, \mathbf{c}_i, \mathbf{f}_i)$, where $\boldsymbol{\mu}_i \in \mathbb{R}^3$ is the 3D mean, $\mathbf{s}_i \in \mathbb{R}^3_+$ the anisotropic scale, $\mathbf{r}_i$ the rotation, o_i the opacity, $\mathbf{c}_i$ the color, and $\mathbf{f}_i$ the language-aligned open-vocabulary feature. Given camera intrinsics $\mathbf{K}_{1:T}$ and globally consistent poses $\mathbf{T}_{1:T}$, we follow [28] to render the Gaussian set $\mathcal{G}$ into each view:

$$(\hat{\mathbf{I}}_t, \hat{\mathbf{D}}_t) \;=\; F(\mathcal{G}, \mathbf{K}_t, \mathbf{T}_t), \qquad t = 1, \ldots, T, \tag{1}$$

where F denotes the differentiable Gaussian rendering operator and $\hat{\mathbf{I}}_t, \hat{\mathbf{D}}_t$ represent render image and depth map. Let $\boldsymbol{\theta}$ collect all Gaussian parameters. The rendered outputs are compared against observations $\{(\mathbf{I}_t, \mathbf{D}_t)\}_{t=1}^T$, and typical 3DGS-based mapping solves

$$\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} \left(\left\| \hat{\mathbf{I}}_t - \mathbf{I}_t \right\|_2^2 + \beta \left\| \hat{\mathbf{D}}_t - \mathbf{D}_t \right\|_2^2 \right), \tag{2}$$

where β balances the RGB and depth terms.

Let $\boldsymbol{\theta}^\star$ denote a solution to Eq. (2). Linearizing the rendering operator around $\boldsymbol{\theta}^\star$ yields $F(\boldsymbol{\theta}^\star + \delta\boldsymbol{\theta}) \approx F(\boldsymbol{\theta}^\star) + \mathbf{J}\,\delta\boldsymbol{\theta}$, where $\mathbf{J} = \frac{\partial F}{\partial \boldsymbol{\theta}}\big|_{\boldsymbol{\theta}^\star}$. For unobservable or weakly observable directions, there exist non-zero perturbations $\delta\boldsymbol{\theta} \neq 0$ such that $\mathbf{J}\,\delta\boldsymbol{\theta} = 0$. Consequently, the solution to Eq. (2) is not isolated, but instead lies on a (near-)manifold in parameter space.

This ambiguity can be further illustrated by considering a single pixel ray $\mathbf{u}$. Under volumetric alpha compositing [44], the rendered color and depth along $\mathbf{u}$ can be written as

$$\hat{\mathbf{I}}(\mathbf{u}) \;=\; \sum_k w_k(\boldsymbol{\theta}; \mathbf{u})\, \mathbf{c}_k, \qquad \hat{\mathbf{D}}(\mathbf{u}) \;=\; \sum_k w_k(\boldsymbol{\theta}; \mathbf{u})\, z_k, \tag{3}$$

where $w_k(\boldsymbol{\theta}; \mathbf{u}) \geq 0$ are ray-dependent compositing weights, and z_k denotes the depth of the k-th contributing Gaussian along the ray. Eq. (3) constrains only the first-order moments along each ray: multiple distinct depth–weight configurations $\{(w_k, z_k)\}$ can yield identical $(\hat{\mathbf{I}}(\mathbf{u}), \hat{\mathbf{D}}(\mathbf{u}))$. Consequently, even with depth supervision, multiple distinct Gaussian configurations $\mathcal{G}$ (equivalently, parameter sets $\boldsymbol{\theta}$) may explain the same observations $\{(\mathbf{I}_t, \mathbf{D}_t)\}_{t=1}^T$. Therefore, minimizing the rendering loss in Eq. (2) does not guarantee a unique or globally consistent 3D geometry, and unconstrained Gaussian updates can gradually erode the geometric consistency provided by the SLAM backend.

Geometrically Anchored Gaussian Updates. To address the above ambiguity, we propose a geometrically consistent 3D Gaussian update strategy with geometry-aware initialization. We parameterize each Gaussian ellipsoid by an anisotropic scale vector $\mathbf{s}$, which controls its principal axis lengths.

For a pixel $\mathbf{u}$ in frame t, we compute the normalized ray direction $\mathbf{d}_{t,\mathbf{u}} \in \mathbb{R}^3$ from intrinsic $\mathbf{K}_t$. We define a local rotation $\mathbf{R}_{t,\mathbf{u}}$ such that its local $+Z$ axis aligns with $\mathbf{d}_{t,\mathbf{u}}$. We initialize a ray-aligned anisotropic scale as

$$\mathbf{s}_{t,\mathbf{u}} = (s_\perp,\, s_\perp,\, s_\parallel), \qquad s_\parallel = \gamma\, s_\perp, \tag{4}$$

where $s_\perp$ and $s_\parallel$ denote the Gaussian extents perpendicular and parallel to the viewing ray, respectively, and γ is a user-controlled elongation ratio. This initialization models each Gaussian as a thin ellipsoid aligned with the sensor ray, providing a geometry-aware prior that reduces ambiguity during optimization. Given SLAM-estimated camera poses $\mathbf{T}_t$ and 3D point positions $\mathcal{P}_t$, Gaussian centers $\boldsymbol{\mu}$ are fixed to $\mathcal{P}_t$, and the optimization problem is formulated as

$$\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} \left(\left\| \hat{\mathbf{I}}_t - \mathbf{I}_t \right\|_2^2 + \beta \left\| \hat{\mathbf{D}}_t - \mathbf{D}_t \right\|_2^2 \right), \text{ s.t. } \boldsymbol{\mu}_t = \mathcal{P}_t \tag{5}$$

C. Open-vocabulary Semantic Association

In contrast to conventional occupancy estimation pipelines that predict *closed-set* semantic classes, a key goal of FreeOcc is to support *open-vocabulary* 3D querying without committing to a fixed label space. To this end, we leverage a pre-trained open-vocabulary segmentation model [57]. Such OV segmentation models produce a *language-aligned* embedding for each pixel, and enable open-vocabulary recognition by computing the similarity between the per-pixel embedding and a text embedding extracted by a language encoder (e.g., CLIP [53]), thereby localizing the region corresponding to an arbitrary textual prompt. Building on this capability, we associate each 3D Gaussian G_i with a language-aligned embedding and construct *language-embedded (LE) Gaussians* [83, 55].

Specifically, given an input image $\mathbf{I}_t$, we extract dense per-pixel embeddings $\mathbf{z}_t(\mathbf{u}) \in \mathbb{R}^D$ at pixel coordinate $\mathbf{u}$ using the OV segmentation model [57]. We then lift these pixel-wise embeddings into 3D using the depth estimated by our SLAM module. For each lifted 3D point, we identify its associated *geometrically anchored* Gaussian in the current Gaussian map and attach the corresponding language-aligned feature to it. The resulting per-Gaussian language features are stored in the Gaussian map, and can be efficiently queried by arbitrary text prompts during the subsequent Gaussian-to-occupancy projection.

D. Gaussian-to-Occupancy Projection

To convert the continuous LE-Gaussian scene representation into a dense volumetric occupancy map, we follow the Gaussian-to-occupancy projection paradigm of GaussianFormer2 [23]. Specifically, we first determine voxel occupancy from the spatial geometry of 3D Gaussian primitives, i.e., their locations and extents in the scene, to estimate whether each voxel is occupied. Different from GaussianFormer2 [23], which targets *closed-set* semantics by aggregating per-class probabilities from Gaussians, we instead construct a *language-embedded* occupancy representation from our LE-Gaussians. Building on the geometric occupancy, we propagate the per-Gaussian language-aligned features into the voxel grid, yielding a *language-embedded occupancy* (LE-occupancy) map that supports open-vocabulary querying.

Concretely, the procedure is as follows. For a query 3D location $\mathbf{x}$, we retrieve its neighboring LE-Gaussian primitives $\mathcal{H}(\mathbf{x}) = \{G_k\}_{k=1}^{P(\mathbf{x})}$, where $P(\mathbf{x}) = |\mathcal{H}(\mathbf{x})|$, and induces the covariance $\boldsymbol{\Sigma}_k = \mathbf{R}(\mathbf{r}_k)\,\mathrm{diag}(\mathbf{s}_k^2)\,\mathbf{R}(\mathbf{r}_k)^\top$, then each neighbor primitive contributes a spatial support

$$\alpha_k(\mathbf{x}) = \exp\!\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right),\ G_k \in \mathcal{H}(\mathbf{x}), \tag{6}$$

and we compose them using probabilistic exclusion:

$$\alpha(\mathbf{x}) = 1 - \prod_{G_k \in \mathcal{H}(\mathbf{x})} \big(1 - \alpha_k(\mathbf{x})\big). \tag{7}$$

To propagate semantics, we compute the posterior responsibility under a local Gaussian mixture model,

$$p(G_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid G_k)\, \pi_k}{\sum_{G_j \in \mathcal{H}(\mathbf{x})} p(\mathbf{x} \mid G_j)\, \pi_j}, \tag{8}$$

where $p(\mathbf{x} \mid G_k) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and π_k is the mixture weight (we set $\pi_k = o_k$, i.e., opacity). We propagate the per-primitive language features to the voxel location by posterior expectation:

$$\mathbf{f}(\mathbf{x}) = \sum_{G_k \in \mathcal{H}(\mathbf{x})} p(G_k \mid \mathbf{x})\, \mathbf{f}_k, \qquad \hat{\mathbf{f}}(\mathbf{x}) = \frac{\mathbf{f}(\mathbf{x})}{\|\mathbf{f}(\mathbf{x})\|_2}. \tag{9}$$

Given a queried category set $\mathcal{C}$, we obtain the corresponding text embeddings $\{\mathbf{t}_c\}_{c \in \mathcal{C}}$ using a text encoder [53] and compute the voxel–text similarity

$$\hat{\mathbf{t}}_c = \frac{\mathbf{t}_c}{\|\mathbf{t}_c\|_2}, \qquad s(\mathbf{x}, c) \;=\; \hat{\mathbf{f}}(\mathbf{x})^\top \hat{\mathbf{t}}_c. \tag{10}$$

This yields an open-vocabulary semantic score at $\mathbf{x}$. We output $\alpha(\mathbf{x})$ and $s(\mathbf{x}, c)$ as the volumetric occupancy probability and open-vocabulary semantic score, respectively. In practice, semantics are reported only for occupied voxels.

V. Experiments

In this section, we seek to answer the following research questions:

- **Q1:** Compared with learning-based occupancy prediction methods, can training-free approaches generalize effectively to unseen datasets? (Sec. V-C)
- **Q2:** Does our Gaussian update strategy offer advantages over well-known 3DGS-SLAM systems in occupancy prediction? How does it impact the FreeOcc system? (Sec. V-D)
- **Q3:** Can our system support online language-based querying of 3D occupancy? (Sec. V-E)

TABLE II: Performance comparison on EmbodiedOcc-ScanNet. Label requirements are reported by task type: **Geo.** indicates required geometric supervision, and **Sem.** indicates semantic supervision. We report IoU and per-class mIoU.

Method	Annotation		IoU	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	mIoU
	Geo.	Sem.													
Fully supervised learning															
TPVFormer [20]	Occupancy	✓	35.88	1.62	30.54	12.03	13.22	35.47	51.39	49.79	25.63	3.6	43.15	16.23	25.70
SurroundOcc [71]	Occupancy	✓	37.04	12.7	31.8	22.5	22	29.9	44.7	36.5	24.6	11.5	34.4	18.2	26.27
GaussianFormer [22]	Occupancy	✓	38.02	17	33.6	21.5	21.7	29.4	47.8	37.1	24.3	15.5	36.2	16.8	27.36
EmbodiedOcc [72]	Occupancy	✓	51.52	22.7	**44.6**	37.4	38	50.1	56.7	59.7	35.4	38.4	52	32.9	42.53
EmbodiedOcc++ [69]	Occupancy	✓	52.2	**27.9**	43.9	38.7	**40.6**	49	**57.9**	59.2	**36.8**	37.8	53.5	34.1	43.60
RoboOcc [79]	Occupancy	✓	**53.3**	21.94	44.57	**39.54**	38.48	**51.28**	57.04	**63.09**	36.7	**43.05**	**54.42**	**34.38**	**44.05**
Self-supervised learning															
GaussianOcc [12]	Poses	✗	10.17	3.81	5.09	2.53	2.56	3.84	10.26	9.9	5.37	0.50	2.33	1.19	4.34
GaussTR [25]	Poses	✗	15.63	1.20	7.78	4.29	2.67	4.52	11.27	10.95	5.31	0.90	4.21	1.34	4.95
Training-free															
Ours (mono)	✗	✗	31.29	3.16	23.49	16.14	13.11	19.66	21.64	23.43	13.76	4.01	8.04	5.98	13.86
Ours (rgbd)	✗	✗	34.40	6.56	26.46	21.69	15.15	21.02	22.09	23.61	15.87	7.48	8.28	6.00	15.84

A. ReplicaOcc Benchmark

Task Definition. We introduce *ReplicaOcc*, a compact, *test-only* benchmark for evaluating *open-vocabulary semantic occupancy prediction* in indoor embodied environments. Similar to EmbodiedOcc-ScanNet [72], incrementally fuse observations over time to maintain globally consistent occupancy prediction. Unlike local frustum-based prediction methods [76], this setting emphasizes long-horizon spatial consistency and semantic completeness, which are essential for embodied agents operating in real environments.

Motivation. Most existing occupancy prediction methods [72, 69, 79] are evaluated exclusively on EmbodiedOcc-ScanNet, which also serves as their primary training source. In many 3D vision tasks (e.g., ACE [3] and LoFTR [61]), training on a single dataset such as ScanNet [7] is often sufficient to generalize to diverse indoor scenes. This motivates the use of a small, test-only dataset to better assess cross-dataset generalization for occupancy prediction.

Dataset Construction. We adopt the Replica [60] sequences release from NICE-SLAM [84]. Following the protocols of Occ-ScanNet [76] and EmbodiedOcc-ScanNet, each scene is converted into a global occupancy grid with resolution $l_x \times l_y \times l_z/(0.08m)^3$, where $l_x \times l_y \times l_z$ denotes the spatial extent of the scene in the world coordinate system. Each voxel is annotated with a binary occupancy label and a semantic category.

B. Experimental Setups

Datasets and Metrics. We evaluate occupancy prediction performance on three datasets: EmbodiedOcc-ScanNet, ReplicaOcc, and EmbodiedOcc-ScanNet-mini. Since each Occ-ScanNet scene clips only 100 frames, it does not satisfy the sequential requirements of SLAM. Therefore, for corresponding scenes, we use monocular or RGB-D sequences from the original ScanNet dataset as SLAM inputs. Evaluation metrics include IoU and mIoU computed on the global scene occupancy, following the EmbodiedOcc evaluation protocol [72]. As fully supervised methods are trained on only 11 semantic categories, mIoU on ReplicaOcc is reported over the 8 categories shared with EmbodiedOcc-ScanNet.

Coordinate System Alignment. During evaluation, inspired by *EVO* [14], SLAM-reconstructed maps are aligned with the occupancy ground-truth coordinate system to resolve global gauge freedom. For monocular or RGB-D SLAM, we estimate a global $\mathrm{Sim}(3) = \{s, \mathbf{R}, \mathbf{t}\}$ or $\mathrm{SE}(3) = \{\mathbf{R}, \mathbf{t}\}$ by aligning camera centers. Given matched SLAM and GT camera poses $\{\mathbf{T}_i^{\text{slam}}, \mathbf{T}_i^{\text{gt}}\}$, we extract camera centers $\mathbf{c}_i^{\text{slam}}, \mathbf{c}_i^{\text{gt}} \in \mathbb{R}^3$ and solve $\min_{s,\mathbf{R},\mathbf{t}} \sum_i \|\mathbf{c}_i^{\text{gt}} - (s\,\mathbf{R}\mathbf{c}_i^{\text{slam}} + \mathbf{t})\|_2^2$ using the closed-form Umeyama solution. The resulting transform is applied consistently to the reconstructed 3D Gaussian map: Gaussian means are updated as $\mathbf{x}' = s\,\mathbf{R}\mathbf{x} + \mathbf{t}$, Gaussian scales as $\boldsymbol{\sigma}' = s\,\boldsymbol{\sigma}$ (or $\log \boldsymbol{\sigma}' = \log \boldsymbol{\sigma} + \log s$), and Gaussian orientations as $\mathbf{R}'_g = \mathbf{R}\mathbf{R}_g$. This alignment step ensures IoU and mIoU computation without affecting the intrinsic reconstruction quality.

C. Occupancy Prediction Evaluation

We evaluate FreeOcc on EmbodiedOcc-ScanNet to assess geometric and semantic occupancy accuracy, and on the proposed ReplicaOcc to examine zero-shot generalization to unseen environments and label spaces.

Baselines. We group baselines by their supervision signals and pose requirements. Results for *fully supervised* methods are taken directly from the corresponding papers [72, 79]. To enable a fair label-free comparison, we additionally implement two *self-supervised learning methods with access to ground-truth camera poses*: GaussianOcc [12] and GaussTR [25]. We focus on Gaussian-based baselines since Gaussian primitives provide a continuous 3D representation and can be naturally fused across frames. Both methods are originally designed for monocular inputs. We adapt them to embodied sequences by fusing per-frame Gaussian predictions into a global coordinate frame using ground-truth poses, followed by a standard Gaussian-to-occupancy conversion [23] to obtain voxelized occupancy maps.

Evaluation on EmbodiedOcc-ScanNet. Quantitative results are reported in Tab. II. As FreeOcc is the first training-free occupancy prediction framework, direct comparisons are limited. Existing self-supervised methods require *ground-*

TABLE III: Zero-shot generalization results on the ReplicaOcc benchmark. The evaluation protocol and categorization follow those in Tab. II.

Method	Annotation		IoU	ceiling	floor	wall	window	chair	bed	sofa	table	mIoU
	Geo.	Sem.										
Fully supervised learning												
EmbodiedOcc [72]	Occupancy	✓	22.91	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.00
Self-supervised learning												
GaussianOcc [12]	Poses	✗	8.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GaussTR [25]	Poses	✗	15.01	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.01
Training-free												
Ours (mono)	✗	✗	**46.81**	**19.38**	**8.88**	**38.23**	**0.21**	**15.11**	**12.14**	**25.07**	**16.46**	**16.93**
Ours (rgbd)	✗	✗	**55.65**	**17.80**	**8.60**	**44.33**	**0.02**	**20.76**	**16.31**	**34.90**	**24.48**	**20.90**

truth camera poses during both training and inference and achieve IoU/mIoU scores of 10.17/4.34 and 15.63/4.95, respectively. In contrast, FreeOcc achieves 31.29/13.86 and 34.40/15.84 IoU/mIoU using monocular and RGB-D inputs, respectively, without any task-specific training. These results exceed self-supervised baselines by more than **2**× across all metrics, demonstrating strong performance without annotation or learned occupancy priors.

Fully supervised methods benefit from dense 3D semantic occupancy annotations, which provide two inherent advantages during evaluation. First, EmbodiedOcc-ScanNet consolidates many object categories into coarse labels such as "objects" and "furniture" [8], introducing ambiguity for open-vocabulary models not explicitly trained on this taxonomy. Second, as shown in Fig. 1, ground-truth labels may deviate from visual evidence, leading to lower IoU/mIoU scores even when predictions better align with real-world observations.

Zero-shot Generalization on ReplicaOcc. To evaluate zero-shot generalization, we directly transfer models trained on EmbodiedOcc-ScanNet to ReplicaOcc without fine-tuning or test-time adaptation. For all methods, per-frame Gaussian primitives are fused into a unified 3D coordinate frame and converted into occupancy volumes using the camera parameters and scene specifications provided by ReplicaOcc.

Quantitative results are reported in Tab. III. FreeOcc exhibits strong zero-shot performance, achieving 46.81/16.93 IoU/mIoU with monocular input and further improving to 55.65/20.90 when RGB-D input is available, substantially outperforming all baselines. In contrast, learning-based occupancy predictors fail to generalize in this transfer setting: their predictions collapse, leading to near-zero per-class scores and mIoU. This degradation can be attributed to two major domain gaps. *(i) Appearance shift:* the scene geometry and visual characteristics of ScanNet differ significantly from those of Replica. *(ii) Camera and scale shift:* although ground-truth poses are used at evaluation time, models trained on ScanNet tend to overfit dataset-specific camera intrinsics and metric scale, which do not transfer reliably across datasets. These results underscore the limited cross-domain generalization of both fully supervised and self-supervised learning-based occupancy predictors, which are prone to overfitting the training distribution and label space. By contrast, our training-free pipeline maintains robust geometric and semantic reasoning across environments without requiring retraining or adaptation.

TABLE IV: Geometric IoU comparison of 3DGS-based SLAM backbones for occupancy prediction on ReplicaOcc and EmbodiedOcc-ScanNet-mini.

Method	IoU		
	Replica	ScanNet-mini	Average
Monocular			
Photo-SLAM [19]	25.03	15.29	20.16
MonoGS [43]	29.50	14.83	22.17
DROID-Splat [16]	26.27	18.41	22.34
Ours (mono)	**46.81**	**31.87**	**39.34**
RGB-D			
SplaTAM [27]	31.11	17.91	24.51
GS-ICP [15]	28.95	21.22	25.09
Photo-SLAM [19]	36.97	16.29	26.63
RTG-SLAM [49]	35.84	18.46	27.15
MonoGS [43]	38.97	19.58	29.28
DROID-Splat [16]	34.48	24.26	29.37
Ours (rgbd)	**55.65**	**34.82**	**45.24**

TABLE V: We report average results of IoU/mIoU and FPS on ReplicaOcc and EmbodiedOcc-ScanNet. **GAGU**: Geometrically Anchored Gaussian Updates. **G-ini**: Geometry-aware initialization.

Method	IoU	mIoU	FPS
Monocular			
w/o GAGU, G-ini	19.88	10.53	10.7
w/o G-ini	31.20	12.06	**26.8**
Ours	**39.05**	**15.40**	25.3
RGB-D			
w/o GAGU, G-ini	27.98	11.20	8.8
w/o G-ini	40.18	16.03	**25.0**
Ours	**45.03**	**18.37**	24.6

Qualitative comparisons are provided in Fig. 3. The above results fully validate that our approach endows occupancy prediction with generalization capabilities, which learning-based methods cannot achieve.

D. Ablation Study

We evaluate the effectiveness of our geometry-consistent Gaussian update strategy and the contributions of its key components (**GAGU**: geometrically anchored Gaussian updates, **G-ini**: geometry-aware initialization) on ReplicaOcc and EmbodiedOcc-ScanNet using monocular and RGB-D inputs.

1) Comparison with 3DGS-SLAM backbones. To assess geometry consistency, we compare FreeOcc against state-of-the-art 3DGS-SLAM systems [19, 43, 16, 27, 15, 49]. For all methods, generated 3D Gaussian maps are aligned using identical procedures (Sec. V-B) and converted into occupancy

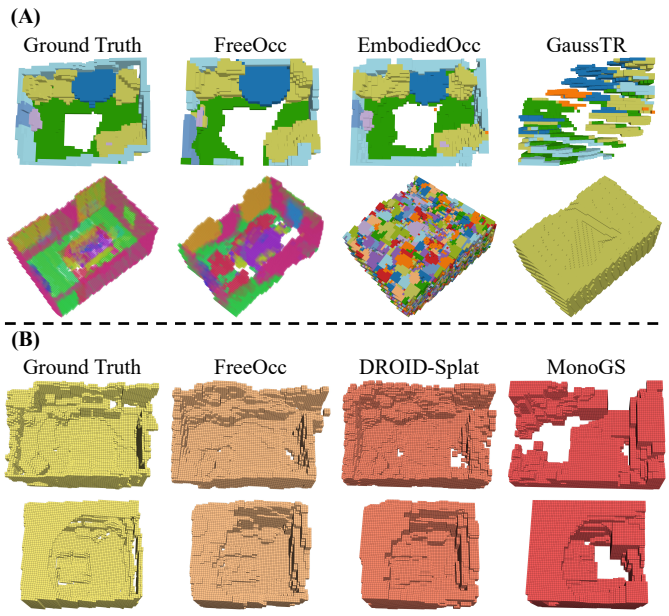

Fig. 3: Qualitative occupancy prediction results. **(A)** Comparisons with learning-based occupancy predictors on "scene0470" and "room2". **(B)** Results of the two 3DGS-SLAM methods with the highest geometric accuracy on "scene0006" and "office0".

volumes (Sec. IV-D) to ensure fair evaluation. As reported in Tab. IV, FreeOcc achieves the highest geometric IoU in both RGB and RGB-D settings, with average improvements of 76.1% and 54.0% over the next-best method (DROID-Splat). The results demonstrate that our geometry-consistent Gaussian update strategy significantly enhances the fidelity of 3D Gaussian construction

2) Component-wise ablation. We further isolate the impact of GAGU and G-ini (Tab. V). Removing GAGU (w/o GAGU, G-ini) significantly reduces IoU/mIoU to 19.88/10.53 (monocular) and 27.98/11.20 (RGB-D) while decreasing FPS. Introducing GAGU alone boosts IoU/mIoU to 31.20/12.06 (monocular) and 40.18/16.03 (RGB-D), improving efficiency by 1.5× and 2.8×. Adding G-ini further increases IoU/mIoU to 39.05/15.40 (monocular) and 45.03/18.37 (RGB-D) with negligible runtime loss. These results confirm that GAGU enforces long-term geometric consistency and accelerates updates, while G-ini enhances initialization for more accurate occupancy reconstruction. Together, they enable robust, real-time, and geometry-consistent 3D Gaussian mapping.

E. Qualitative Results

We present qualitative visualizations corresponding to the quantitative evaluations in Sec. V-C and Sec. V-D. Representative results are shown in Fig. 3. Fig. 3(A) compares FreeOcc with learning-based occupancy prediction methods on EmbodiedOcc-ScanNet and ReplicaOcc scenes. While supervised and self-supervised methods produce incomplete or near-empty occupancy maps on ReplicaOcc, FreeOcc consistently reconstructs coherent geometric structures with meaningful semantic occupancy across datasets. These results visually corroborate the strong zero-shot generalization performance

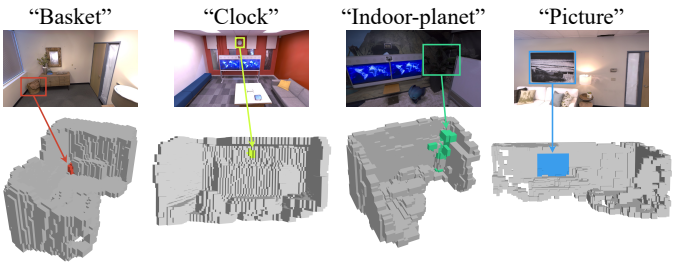

Fig. 4: Open-vocabulary query results on ReplicaOcc, demonstrating semantic occupancy retrieval for different input vocabulary words.

observed in our quantitative analysis. Fig. 3(B) compares FreeOcc against the two 3DGS-SLAM methods with the highest geometric accuracy. Despite using similar SLAM backbones, our geometry-consistent Gaussian update strategy yields more complete and spatially consistent occupancy maps, particularly around object boundaries and thin structures, highlighting the benefits of tightly coupling Gaussian refinement with SLAM geometry.

We further visualize open-vocabulary querying results on ReplicaOcc in Fig. 4. We evaluate challenging queries involving small objects (e.g., "basket" and "clock"), low-light environments (e.g., "indoor planet"), and semantically ambiguous categories (e.g., "picture"). FreeOcc successfully localizes and retrieves these targets directly from the occupancy map, demonstrating robust open-vocabulary semantic grounding in online occupancy prediction.

VI. Conclusion

This work addresses the challenge of open-vocabulary occupancy prediction, where existing methods rely heavily on dense geometric and semantic annotations and often generalize poorly to novel environments. We introduced FreeOcc, the first training-free framework for open-vocabulary occupancy prediction in embodied indoor settings. By designing a multi-layer mapping pipeline that tightly integrates SLAM geometry, 3D Gaussian representations, and vision–language semantics, FreeOcc constructs globally consistent occupancy maps without task-specific training or voxel-level supervision. Extensive experiments demonstrate that FreeOcc achieves competitive geometric and semantic accuracy while significantly improving generalization compared to learning-based approaches. Despite its strong performance, FreeOcc still depends on the robustness of the SLAM backbone, and its semantic association may be affected by temporally inconsistent VLM predictions. Future work will explore occupancy-derived geometric and semantic cues for SLAM optimization, as well as confidence-aware temporal filtering, to further improve long-term consistency and robust open-vocabulary 3D scene understanding. We believe this work represents an important step toward scalable, annotation-free occupancy prediction, enabling robots to reason about arbitrary 3D environments and semantics in real-world deployments.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No.62573370), the Key Area Project of Education Department of Guangdong Province (No.2025ZDZX3051) and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007).

REFERENCES

[1] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Langocc: Open vocabulary occupancy estimation via volume rendering. In *International Conference on 3D Vision 2025*, 2025.

[2] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Gaussianflowocc: Sparse and weakly supervised occupancy estimation using gaussian splatting and temporal flow, 2025.

[3] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[4] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2017.

[5] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022.

[6] Lih-Shyang Chen and Marc R Sontag. Representation, display, and manipulation of 3d digital scenes and their medical applications. *Computer Vision, Graphics, and Image Processing*, 48(2):190–216, 1989.

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[9] Tianchen Deng, Yue Pan, Shenghai Yuan, Dong Li, Chen Wang, Mingrui Li, Long Chen, Lihua Xie, Danwei Wang, Jingchuan Wang, Javier Civera, Hesheng Wang, and Weidong Chen. What is the best 3d scene representation for robotics? from geometric to foundation models, 2026.

[10] Ben Fei, Jingyi Xu, Rui Zhang, Qingyuan Zhou, Weidong Yang, and Ying He. 3d gaussian splatting as new era: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[11] Gideon Frieder, Dan Gordon, and R Reynolds. Back-to-front display of voxel based objects. *IEEE Computer Graphics and Applications*, 5(01):52–60, 1985.

[12] Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[13] Yuhang Gao, Xiang Xiang, Sheng Zhong, and Guoyou Wang. Loc: A general language-guided framework for open-set 3d occupancy prediction. *arXiv preprint arXiv:2510.22141*, 2025.

[14] Michael Grupp. evo: Python package for the evaluation of odometry and slam. https://github.com/MichaelGrupp/evo, 2017.

[15] Seongbo Ha, Jiung Yeon, and Hyeonwoo Yu. Rgbd gs-icp slam. In *European Conference on Computer Vision*, page 180–197. Springer, 2020.

[16] Christian Homeyer, Leon Begiristain, and Christoph Schnörr. Droid-splat combining end-to-end slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

[17] Armin Hornung, Kai M. Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots*, 2013. doi: 10.1007/s10514-012-9321-0.

[18] Yan Song Hu, Nicolas Abboud, Muhammad Qasim Ali, Adam Srebrnjak Yang, Imad Elhajj, Daniel Asmar, Yuhao Chen, and John S. Zelek. Mgso: Monocular real-time photometric slam with efficient 3d gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11061–11067, 2025. doi: 10.1109/ICRA55743.2025.11127380.

[19] Huajian Huang, Longwei Li, Hui Cheng, and Sai-Kit Yeung. Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular stereo and rgb-d cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[20] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023.

[21] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[22] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *European Conference on Computer Vision*, pages 376–393. Springer, 2024.

[23] Yuanhui Huang, Amonnut Thammatadatrakoon, Wen-

zhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27477–27486, 2025.

[24] Yiming Ji, Yang Liu, Guanghu Xie, Boyu Ma, Zongwu Xie, and Hong Liu. Neds-slam: A neural explicit dense semantic slam framework using 3d gaussian splatting. *IEEE Robotics and Automation Letters*, 9(10): 8778–8785, October 2024. ISSN 2377-3774. doi: 10.1109/lra.2024.3451390.

[25] Haoyi Jiang, Liu Liu, Tianheng Cheng, Xinjie Wang, Tianwei Lin, Zhizhong Su, Wenyu Liu, and Xinggang Wang. Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[26] Haoyi Jiang, Liu Liu, Tianheng Cheng, Xinjie Wang, Tianwei Lin, Zhizhong Su, Wenyu Liu, and Xinggang Wang. Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[27] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[28] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.

[30] Xiaolei Lang, Laijian Li, Chenming Wu, Chen Zhao, Lina Liu, Yong Liu, Jiajun Lv, and Xingxing Zuo. Gaussian-lic: Real-time photo-realistic slam with gaussian splatting and lidar-inertial-camera fusion. In *2025 International Conference on Robotics and Automation*. IEEE, 2025.

[31] Xiaolei Lang, Jiajun Lv, Kai Tang, Laijian Li, Jianxin Huang, Lina Liu, Yong Liu, and Xingxing Zuo. Gaussian-lic2: Lidar-inertial-camera gaussian splatting slam. *arXiv preprint arXiv:2507.04004*, 2025.

[32] Bernard Lange, Masha Itkina, Jiachen Li, and Mykel Kochenderfer. Self-supervised multi-future occupancy forecasting for autonomous driving. *Robotics: Science and Systems*, 2025.

[33] Haoang Li, Xiangqi Meng, Xingxing Zuo, Zhe Liu, Hesheng Wang, and Daniel Cremers. Pg-slam: Photo-realistic and geometry-aware rgb-d slam in dynamic environments. *IEEE Transactions on Robotics*, 2025.

[34] Linfei Li, Lin Zhang, Zhong Wang, and Ying Shen. Gs3lam: Gaussian semantic splatting slam. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 3019–3027, New York, NY, USA, 2024. Association for Computing Machinery.

[35] Mingrui Li, Shuhong Liu, Heng Zhou, Guohao Zhu, Na Cheng, Tianchen Deng, and Hongyu Wang. Sgs-slam: Semantic gaussian splatting for neural dense slam. In *European Conference on Computer Vision*, pages 163–179. Springer, 2024.

[36] Peizheng Li, Shuxiao Ding, You Zhou, Qingwen Zhang, Onat Inak, Larissa Triess, Niklas Hanselmann, Marius Cordts, and Andreas Zell. Ago: Adaptive grounding for open world 3d occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8645–8655, October 2025.

[37] Xiang Li, Yupeng Zheng, Pengfei Li, Yilun Chen, Ya-Qin Zhang, and Wenchao Ding. Enhancing indoor occupancy prediction via sparse query-based multi-level consistent knowledge distillation. *IEEE Robotics and Automation Letters*, 2025.

[38] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023.

[39] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. FB-OCC: 3D occupancy prediction based on forward-backward view transformation. *arXiv:2307.01492*, 2023.

[40] Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. Octreeocc: Efficient and multi-granularity occupancy prediction using octree queries. *Advances in Neural Information Processing Systems*, 37:79618–79641, 2024.

[41] Jiajun Lv, Kewei Hu, Jinhong Xu, Yong Liu, Xiushui Ma, and Xingxing Zuo. Clins: Continuous-time trajectory estimation for lidar-inertial system. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 6657–6663. IEEE, 2021.

[42] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *Advances in Neural Information Processing Systems*, 39, 2025.

[43] Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian Splatting SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[44] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, December 2021. ISSN 0001-0782. doi: 10.1145/3503250.

[45] Raúl Mur-Artal and Juan D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Transactions on Robotics*, 33 (5):1255–1262, 2017. doi: 10.1109/TRO.2017.2705103.

[46] Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[47] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation*, pages 12404–12411. IEEE, 2024.

[48] Yue Pan, Xingguang Zhong, Liren Jin, Louis Wiesmann, Marija Popovic, Jens Behley, and Cyrill Stachniss. Pings: Gaussian splatting meets distance fields within a point-based implicit neural map. *Robotics: Science and Systems*, 2025.

[49] Zhexi Peng, Tianjia Shao, Liu Yong, Jingke Zhou, Yin Yang, Jingdong Wang, and Kun Zhou. Rtg-slam: Real-time 3d reconstruction at scale using gaussian splatting. *ACM SIGGRAPH Conference Proceedings, Denver, CO, United States, July 28 - August 1, 2024*, 2024.

[50] Mason Peterson, Yixuan Jia, Yulun Tian, Annika Thomas, and Jonathan P. How. Roman: Open-set object map alignment for robust view-invariant global localization. *Robotics: Science and Systems*, 2025.

[51] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pages 194–210. Springer, 2020.

[52] Matteo Poggi, Fabio Tosi, Fatma Güney, and Sadra Safadoust. Self-evolving depth-supervised 3d gaussian splatting from rendered stereo pairs, 2024.

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[54] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[55] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 2024.

[56] Yiang Shi, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Xinggang Wang. Occupancy as set of points. In *European Conference on Computer Vision*, pages 72–87. Springer, 2024.

[57] Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. *arXiv preprint arXiv:2411.09219*, 2024.

[58] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *arXiv preprint arXiv:1611.08974*, 2016.

[59] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017.

[60] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[61] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[62] Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15035–15044, 2024.

[63] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020.

[64] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in neural information processing systems*, 2021.

[65] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979.

[66] Evangelos Ververas, Rolandos Alexandros Potamias, Jifei Song, Jiankang Deng, and Stefanos Zafeiriou. Sags: Structure-aware 3d gaussian splatting. In *European Conference on Computer Vision*, pages 221–238. Springer, 2024.

[67] Antonin Vobecky, Oriane Siméoni, David Hurych, Spyridon Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Pop-3d: Open-vocabulary 3d occupancy prediction from images. *Advances in Neural Information Processing Systems*, 36:50545–50557, 2023.

[68] Fengyun Wang, Dong Zhang, Hanwang Zhang, Jinhui Tang, and Qianru Sun. Semantic scene completion with cleaner self. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–877, 2023.

[69] 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 *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 925–934, New York, NY, USA, 2025. Association for Computing Machinery.

[70] Ruoyu Wang, Yukai Ma, Yi Yao, Sheng Tao, Haoang Li, Zongzhi Zhu, Yong Liu, and Xingxing Zuo. L2cocc: Lightweight camera-centric semantic scene completion via distillation of lidar model. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 716–723. IEEE, 2025.

[71] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023.

[72] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 26360–26370, October 2025.

[73] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[74] Dianyi Yang, Yu Gao, Xihan Wang, Yufeng Yue, Yi Yang, and Mengyin Fu. Opengs-slam: Open-set dense semantic slam with 3d gaussian splatting for object-level scene understanding, 2025.

[75] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *2023 IEEE/CVF International Conference on Computer Vision*, pages 9421–9431, 2023.

[76] Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In *European Conference on Computer Vision*, pages 38–54. Springer, 2024.

[77] Vladimir Yugay, Yue Li, Theo Gevers, and Martin R. Oswald. Gaussian-slam: Photo-realistic dense slam with gaussian splatting, 2023.

[78] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *CoRR*, abs/2312.09243, 2023.

[79] 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*, 2025.

[80] Jilai Zheng, Pin Tang, Zhongdao Wang, Guoqing Wang, Xiangxuan Ren, Bailan Feng, and Chao Ma. Veon: Vocabulary-enhanced occupancy prediction. In *European Conference on Computer Vision*, pages 92–108. Springer, 2024.

[81] Changqing Zhou, Yueru Luo, and Changhao Chen. Generalizing visual geometry priors to sparse gaussian occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[82] Changqing Zhou, Yueru Luo, Han Zhang, Zeyu Jiang, and Changhao Chen. Monocular open vocabulary occupancy prediction for indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[83] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024.

[84] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.