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

SC-VLMaps: Depth-Free Visual–Language Mapping via Scene Coordinate Regression

Nanda Febri Istighfarin[1], Baehoon Choi[2], and HyungGi Jo[1,*]

***Abstract*— The ability to connect visual observations with human language is increasingly valuable for embodied agents in tasks such as navigation and semantic mapping. Existing visual–language map (VLMaps) approach enables this connection but typically depends on depth images to project semantic features into 3D space, which limits scalability due to sensor cost and deployment constraints. In this work, we introduce SC-VLMaps, a depth-free visual–language mapping framework that constructs semantic maps using only monocular RGB input. SC-VLMaps leverages a scene coordinate regression (SCR) network to predict dense 3D coordinates from images, bypassing the need for depth supervision and enabling implicit geometry reconstruction. The predicted coordinates are fused into a voxel grid and augmented with language-aligned features from a frozen visual–language encoder, producing maps that are both geometrically coherent and semantically enriched. By employing a multi-scene training strategy, SC-VLMaps generalizes from indoor datasets (7Scenes) to challenging outdoor benchmarks (Cambridge Landmarks). Experiments show that SC-VLMaps achieves denser, more compact maps with stronger semantic alignment than VLMaps, while requiring only monocular RGB images.**

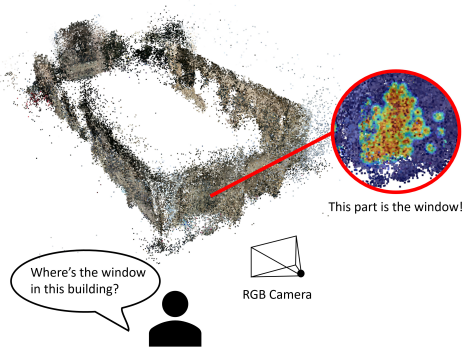

Fig. 1. SC-VLMaps is a depth-free visual–language mapping framework that constructs spatial maps using only RGB input. A scene coordinate regression (SCR) network predicts dense 3D coordinates from monocular images, which are fused into a voxel grid and augmented with language-aligned features. This enables natural-language indexing and queries over both indoor and outdoor environments, without requiring depth sensors or explicit 3D reconstruction.

I. INTRODUCTION

Humans possess an extraordinary ability to interpret and navigate their surroundings. This capability is based on cognitive maps, which capture where objects are, how they relate to each other, and how to navigate between them. For decades, robotics research has aimed to support machines with similar abilities, enabling robots to build and use maps for perception, reasoning, and action. Among the many approaches, one promising direction is to enrich maps with semantic and linguistic information, so that they can be indexed and queried directly through human natural language. Such visual–language maps enable new forms of human–robot interaction and support map-centric tasks, including navigation, exploration, and simultaneous localization and mapping (SLAM).

The concept of a visual-language map for robotics was formally introduced by VLMaps [1]. VLMaps demonstrated how to embed pretrained visual-language features [2] into a 3D spatial map, making it possible to ground open-vocabulary natural language queries in physical environments. Prior to this, related efforts existed but were fragmented: language-grounded navigation tasks [3], semantic mapping [4], and language-conditioned task execution [5]. While these works advanced the field, they did not unify language, vision, and spatial mapping into a single representation. VLMaps filled this gap, but at the cost of requiring depth sensors to explicitly reconstruct point clouds. This reliance on depth data remains a limitation, especially as many practical robotic platforms favor cheaper monocular RGB cameras.

In this work, we extend the idea of visual-language maps by eliminating the dependence on depth sensors. Instead of constructing explicit point clouds from depth images, we leverage implicit 3D scene coordinate maps predicted by a scene coordinate regression (SCR) network. Our approach enables the creation of visual-language maps from RGB images alone. Furthermore, the pretrained SCR model generalizes across data splits: once trained on a dataset (e.g., the training split of Cambridge Landmarks), it can be reused to generate maps for unseen scenes or test sequences without retraining, making the framework scalable and efficient. Finally, while VLMaps was originally evaluated only on indoor datasets, we expand its scope to both indoor and outdoor settings.

Accordingly, our contributions are summarized as follows:

- We propose SC-VLMaps, an RGB-only variant of VLMaps that incorporates implicit 3D scene coordinate maps, thereby eliminating the need for depth sensors.
- We show that a pretrained SCR network generalizes

※ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00346415)

[1]Nanda Febri Istighfarin and HyungGi Jo are with the Division of Electronic Engineering, Jeonbuk National University, Jeonju 54896, Korea (e-mails: nnd.fbr, hygijo@jbnu.ac.kr)

[2]Baehoon Choi is with Bstar Robotics Co., Ltd. (e-mail: baehoon@bstar.tech)

*Corresponding author.

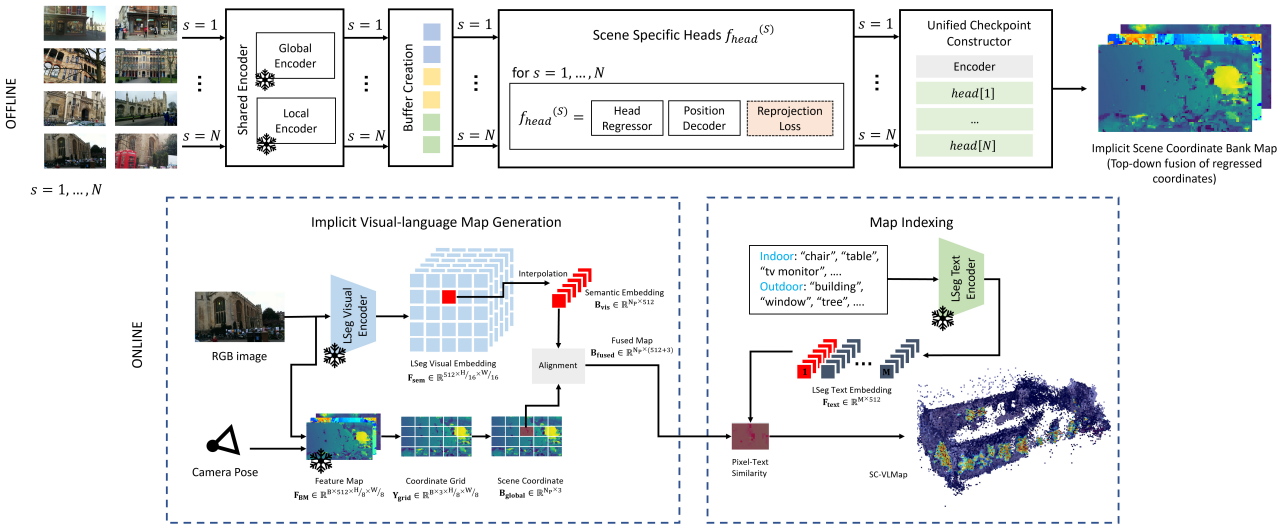

Fig. 2. Overview of the proposed framework. The process is divided into two stages: offline and online. The offline stage focuses on training the implicit scene coordinate bank, while the online stage is responsible for constructing the visual–language map.

across dataset splits, enabling map construction on unseen sequences without additional training.

- We extend the scope of visual–language maps from indoor to outdoor environments, and validate our approach on both the 7Scenes dataset and Cambridge Landmarks.

II. RELATED WORKS

Semantic Mapping. Semantic mapping has evolved through successive stages. Early systems used RGB-D input to assign class labels to 3D maps [4]. Later work shifted from broad classes to instance-level and moving-object tracking [6], even introducing structured maps that represent places, rooms, objects, and their relationships [7]–[9]. More recently, language-aligned visual features have been fused into 3D [1], enabling robots to query and navigate using natural language. At the same time, sensing has shifted from depth cameras toward RGB-only setups. This shift motivates our approach: building visual–language maps from RGB alone, without depth.

Visual-language Model. Vision–language models have gained increasing attention over the past decade. The line of work starts with [10], which projects images into a word-embedding space, and then grows to transformer-based pretraining with region features, such as [11]. The big shift comes with CLIP [12], which introduces a large-scale, contrastive framework that aligns visual and textual representations in a shared embedding space. Originally aimed at image–text retrieval and classification, CLIP encodes semantic concepts as text embeddings, enabling flexible, prompt-driven understanding of visual content. These ideas have since been extended to semantic tasks [2], [13].

Scene Coordinate Regression (SCR). SCR performs visual localization by using a neural network to predict a 3D scene coordinate for every image pixel, producing 2D–3D correspondences without the explicit keypoint detection, description, or matching required by feature-based methods. Early SCR variants [14], [15] were limited by long training times and poor scalability to large scenes. The recent SCR approach [16] that introduces gradient decorrelation substantially accelerates training while also improving pose accuracy. Building on this, the extended variant [17] incorporates additional modules designed to prioritize more informative features during regression. More recently, GLACE [18] augments local features with global scene encodings, providing broader context to the regressor and enabling SCR to scale to larger, more complex environments with stronger localization performance.

III. METHOD

A. Overview

Our objective is to build a visual-language map from RGB images without relying on depth. Our proposed framework is summarized in Fig. 2. Instead of reconstructing a traditional point cloud or mesh, our method predicts dense 3D scene coordinates for each image and fuses them across multiple views to form an implicit map. This geometric representation is further enriched with semantic features extracted from a vision-language model, enabling text-driven map interaction.

As shown in Fig. 2, our method consists of two primary stages:

- an offline scene coordinate regression (SCR) training phase, where a shared encoder and scene-specific heads are optimized to predict 3D coordinates; and
- an online mapping phase, where these predictions are fused with language-aligned embeddings to generate a top-down, semantically indexed map.

We discussed the offline stage in Section III-B, while the online stage is covered in Sections III-C.

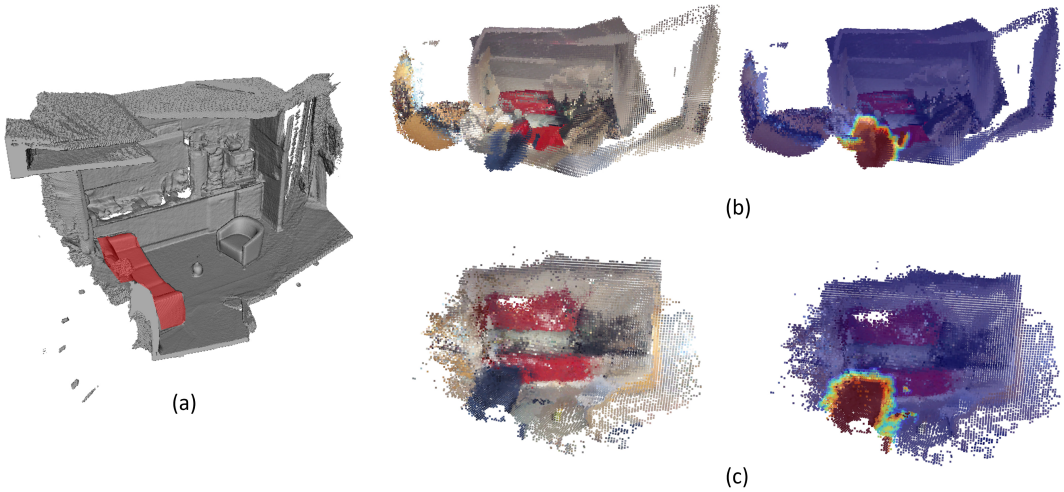

Fig. 3. Comparison of map generation results on the 7Scenes Pumpkin dataset. (a) Illustrates the ground-truth mesh, with the red region indicating the sofa location. (b) Presents the map produced by VLMaps together with its predicted sofa location. (c) Shows the map generated by our method along with the corresponding sofa prediction.

B. Multi-Scene Coordinate Regression

Scene coordinate regression (SCR) offers a learning-based alternative to traditional feature-matching pipelines for visual localization. Like feature-based approaches, SCR consists of two stages: (1) generating 2D–3D correspondences and (2) estimating the camera pose. In this work, we focus solely on the first stage—predicting dense 2D–3D correspondences via a neural network—while assuming ground-truth poses are available.

We adopt the GLACE framework [18], which builds upon ACE [16] to scale SCR to large scenes without requiring ground-truth 3D supervision. Given an RGB image I, the network outputs a dense map of 3D scene coordinates $y_k \in \mathbb{R}^3$, one per pixel x_k. These coordinates are learned through reprojection supervision, leveraging known camera poses and intrinsics.

Unlike monocular depth estimation methods that regress per-pixel camera-frame depth, SCR predicts 3D scene coordinates directly in a global reference frame. The network does not estimate camera-centric depth maps. Instead, the metric scale emerges from reprojection consistency with known camera poses and intrinsics. Consequently, the model learns scene-level geometry without requiring depth maps or external metric depth datasets.

Feature Encoding and Architecture. To handle large environments with spatially ambiguous patterns, we adopt a dual-branch feature encoder:

- A local encoder (DSAC* [15]) captures high-resolution, patch-level details.
- A global encoder (R2Former [19]) provides scene-wide context.

The resulting features are concatenated and passed through a scene-specific head, which regresses the 3D coordinate for each pixel. To reduce training cost, features are precomputed and stored in a training buffer, allowing efficient sampling without redundant forward passes.

Reprojection-based Supervision. Our training does not require ground-truth 3D coordinates. Instead, we exploit reprojection consistency [15], [16], [20]. For each pixel $x_k \in \mathbb{R}^2$ in the image, the network predicts a 3D scene coordinate $y_k \in \mathbb{R}^3$. Using the ground-truth camera intrinsics $\mathbf{K}$ and pose $[\mathbf{R}^*|\mathbf{t}^*]$, this coordinate is projected back into the image plane. The reprojection error is defined as:

$$r(x_k, y_k) = \left\| x_k - \pi\left(\mathbf{K}(\mathbf{R}^* y_k + \mathbf{t}^*)\right) \right\|_2, \tag{1}$$

where $\pi(\cdot)$ denotes the projection from 3D to 2D.

To improve robustness, the reprojection error is passed through a dynamic tanh-based robust loss:

$$\ell(x_k, y_k) = \tau(t) \cdot \tanh\left(\frac{r(x_k, y_k)}{\tau(t)}\right), \tag{2}$$

where the bandwidth $\tau(t)$ decreases over normalized training time $t \in (0, 1)$ as:

$$\tau(t) = \sqrt{1 - t^2} \cdot \tau_{\text{max}} + \tau_{\text{min}} \tag{3}$$

A prediction is considered valid if its depth falls between 10 cm and 1000 m, and its reprojection error is less than 1000 pixels. Otherwise, a fallback regularization is applied:

$$\ell(x_k, y_k) = \left\| y_k - \hat{y}_k \right\|_2, \tag{4}$$

where $\hat{y}_k = \mathbf{R}^* \hat{e}_k + \mathbf{t}^*$ is a synthetic point placed at a fixed depth (e.g., 10 meters) along the ray corresponding to pixel x_k. This regularizer helps stabilize gradients early in training.

Multi-Scene Training Protocol. To scale across diverse environments, we adopt a multi-scene training strategy. The network consists of a shared encoder and a set of scene-specific regression heads. For an input from scene s, the shared encoder extracts features, which are passed to the corresponding head $f_{\text{head}}^{(s)}$ for coordinate regression.

TABLE I

COMPARISON OF MAP STATISTICS ACROSS THE 7SCENES INDOOR DATASETS. HIGHER DENSITY (↑) REFLECTS IMPROVED MAP QUALITY, VOXEL COUNTS (∼) AND SCENE VOLUME [m^3] ARE REPORTED FOR REFERENCE. **BOLD** DENOTES THE BEST RESULTS.

	Chess	Fire	Heads	Office	Pumpkin	Red Kitchen	Stairs
Density (↑)							
VLMaps	0.000786	0.000382	0.000135	0.001706	0.001096	0.001448	0.000618
Ours (SC-VLMaps)	**0.001109**	**0.003439**	**0.001843**	**0.008398**	**0.003348**	**0.007403**	**0.000648**
Occupied Voxels (∼)							
VLMaps	101,902	47,426	13,311	230,512	154,822	218,041	81,135
Ours (SC-VLMaps)	50,151	38,525	6,521	58,570	46,540	37,794	30,122
Scene Volume [m^3]							
VLMaps	1,220.97	404.09	442.38	871.79	1,737.63	1,044.61	19,622.79
Ours (SC-VLMaps)	1,220.97	404.09	442.38	871.79	1,737.63	1,044.61	19,622.79

Training is conducted episodically: each mini-batch contains samples from only one scene, and updates only the matching head while allowing the encoder to accumulate gradients across all scenes. This avoids inter-scene interference while maintaining shared representation learning.

The final model is serialized as a unified checkpoint, bundling the encoder and all trained heads. In the online stage, the correct head is selected via scene ID, enabling accurate coordinate regression across multiple environments using a compact model.

C. Implicit Visual-Language Map Construction

The trained SCR model enables per-frame prediction of 3D coordinates for each pixel. To construct a globally consistent and semantically indexed map, we fuse these predictions across frames and augment them with language-aligned descriptors.

Map Generation. Given an input image and its camera pose, we infer the dense scene coordinate map $Y_{\text{grid}} \in \mathbb{R}^{3\times H/8\times W/8}$, and transform it into the global coordinate frame:

$$B_{\text{global}} = \mathbf{R} \cdot Y_{\text{grid}} + \mathbf{t}. \tag{5}$$

Then, flattening yields $B_{\text{global}} \in \mathbb{R}^{N_p\times 3}$, where $N_p = (H/8)\cdot(W/8)$. Across time, multiple such coordinate maps are aggregated into a top-down voxel grid. Each voxel cell accumulates the 3D coordinates projected into its space, and averages them to produce a geometrically consistent implicit map.

To embed semantic meaning into the map, we leverage the frozen LSeg encoder. The RGB image is processed into a dense semantic feature map $F_{\text{sem}} \in \mathbb{R}^{512\times H/16\times W/16}$. For each coordinate in B_{global}, we interpolate the semantic feature at its projected 2D location, obtaining a visual-language embedding $B_{\text{vis}} \in \mathbb{R}^{N_p\times 512}$.

We then construct a fused representation:

$$B_{\text{fused}} = \left[B_{\text{global}} \parallel B_{\text{vis}}\right], \tag{6}$$

so that $B_{\text{fused}} \in \mathbb{R}^{N_p\times 515}$. These fused descriptors are pooled within each voxel cell, yielding a final representation that couples geometry with semantic content.

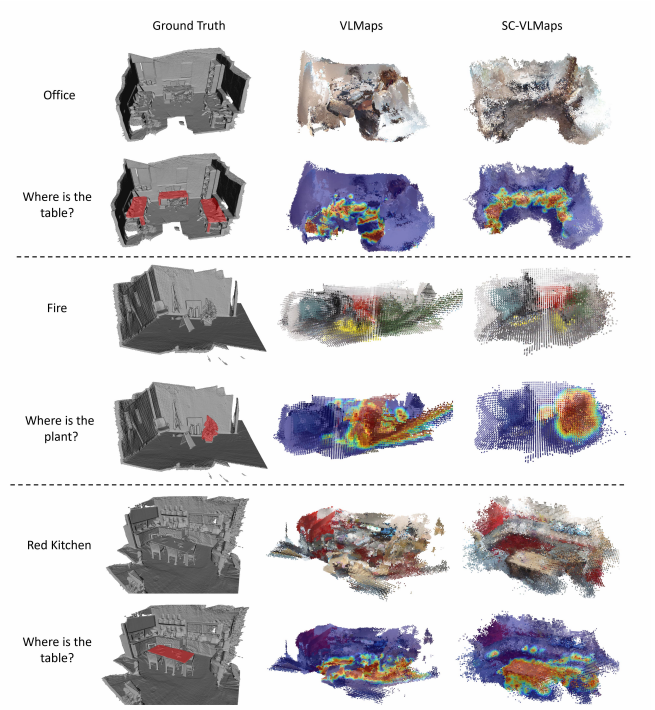

Fig. 4. Qualitative comparison of map generation and prompt-based retrieval in 7Scenes. Each row shows the ground-truth reconstruction, the VLMaps baseline, and our proposed SC-VLMaps. For Office and Red Kitchen, we query "Where is the table?", while for Fire we query "Where is the plant?". SC-VLMaps produces denser reconstructions and more accurate localization of the queried objects compared to VLMaps.

Map Indexing. To enable natural-language querying, we embed a predefined prompt set using the LSeg text encoder. Prompts are domain-specific, covering both indoor and outdoor categories (e.g., "table", "building", "sky", etc.), producing text embeddings $F_{\text{text}} \in \mathbb{R}^{M\times 512}$.

For each voxel, we compute the cosine similarity between its semantic descriptor and each prompt embedding, resulting in pixel-text similarity maps that encode textual relevance per location.

This allows interactive exploration of the map via text queries, e.g., "where are the windows?", "show tree regions", etc., without explicit segmentation or classification.

While prior work in visual-language mapping has largely

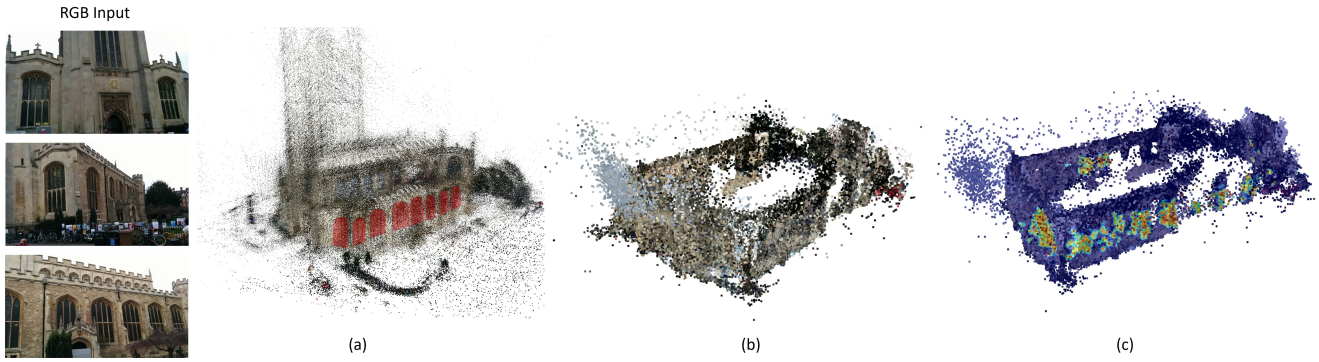

Fig. 5. Reconstruction results on the Cambridge Landmarks St. Mary's Church dataset. The RGB inputs used for reconstruction are shown on the left. (a) Ground-truth reconstruction, where the red regions highlight the presence of windows. (b) Reconstructed point cloud generated by our SC-VLMaps approach. (c) Corresponding semantic heatmap for the prompt"window"

TABLE II

SEMANTIC ALIGNMENT COMPARISON ON THE 7SCENES DATASET. REPORTED METRICS INCLUDE THE SEMANTIC ALIGNMENT SCORE (SAS) AND COVERAGE AT THRESHOLDS OF 70, 80, AND 90, WHERE HIGHER VALUES INDICATE BETTER PERFORMANCE. BOLD VALUES DENOTE THE BEST RESULT BETWEEN THE COMPARED METHODS.

Scene	Ours (SC-VLMaps)				VLMaps			
	SAS	Cov@70	Cov@80	Cov@90	SAS	Cov@70	Cov@80	Cov@90
Chess	**0.912**	1.000	0.999	**0.686**	0.900	1.000	**1.000**	0.519
Fire	0.850	0.997	0.791	**0.253**	**0.857**	0.997	**0.884**	0.171
Heads	**0.894**	1.000	**0.976**	**0.520**	0.880	1.000	0.975	0.295
Office	**0.887**	1.000	0.933	**0.485**	0.885	1.000	**0.993**	0.397
Pumpkin	**0.884**	1.000	**0.997**	**0.344**	0.868	1.000	0.995	0.166
Red Kitchen	**0.899**	0.999	**0.981**	**0.541**	0.886	**1.000**	0.995	0.358
Stairs	**0.936**	1.000	**0.998**	**0.743**	0.926	1.000	0.995	0.591

focused on indoor environments, our formulation naturally extends to outdoor domains. The use of LSeg ensures robustness to unstructured categories such as sky, vegetation, or glass. This is particularly important in urban or campus-scale scenarios, where structured landmarks are sparse and semantic variation is high.

IV. EXPERIMENTS

A. Experiment Setup

Datasets. We evaluate our approach on two publicly available benchmark datasets: 7Scenes [14] and Cambridge Landmarks [21]. The 7Scenes dataset provides indoor sequences with cluttered layouts and repeated textures with a relatively small range. Cambridge Landmarks extends the evaluation to outdoor urban environments. This setting is far more challenging: scenes span much larger spatial extents, lighting conditions change dramatically across captures, and many structures (e.g., glass facades, repetitive windows, vegetation, or open sky) are difficult to model reliably. While prior visual–language mapping methods such as VLMaps have been limited to depth-equipped indoor datasets, our formulation allows us to address both indoor and outdoor domains directly from RGB images.

Training and Evaluation Setup. The proposed method extends the GLACE framework [18] by incorporating a multi-scene regressor composed of a shared encoder and scene-specific regression heads. This architecture facilitates the learning of scene-invariant representations while retaining specialization for individual environments.

Training is conducted on the training split of each dataset, using only reprojection-based supervision, without access to ground-truth 3D coordinates. The model is trained using RGB images and known camera poses, and does not rely on depth maps at any stage. All input images are converted to grayscale and resized such that the shorter side is 480 pixels. Please note that the offline stage is trained on an NVIDIA RTX 6000 Ada Generation GPU (48 GB memory), ensuring sufficient capacity for large-scale multi-scene learning.

During the online stage, the coordinate regression network is kept fixed, and maps are constructed solely from the test sequences on a NVIDIA RTX 4070 Ti GPU (12 GB memory). Average end-to-end processing time during mapping is approximately 250–266 ms per frame (≈3.8–4.0 FPS). Scene coordinate prediction itself requires about 5 ms per frame, with the majority of computation attributed to semantic feature extraction. This ensures that online map generation uses unseen images, providing a fair comparison with prior methods.

In the mapping stage, the predicted scene coordinates are fused into a voxel grid with varying resolution depending on the dataset. Each voxel is augmented with semantic descriptors obtained from a frozen LSeg encoder [2], enabling language-conditioned queries over the reconstructed map.

As a baseline, VLMaps [1] is re-implemented by combin-

TABLE III

TOP-5 SEMANTIC CATEGORIES PER SCENE. WE COMPARE SC-VLMAPS WITH VLMAPS, REPORTING THE FRACTION OF OCCUPIED VOXELS ASSOCIATED WITH EACH CATEGORY.

Scene	Ours (SC-VLMaps)	VLMaps
Chess	Floor(0.24), Table(0.19), Picture(0.16), Wall(0.15), Chair(0.10)	Wall(0.29), Table(0.20), TV(0.15), Chair(0.10), Picture(0.10)
Fire	Floor(0.35), Wall(0.21), Picture(0.18), Plant(0.17), Chair(0.08)	Wall(0.26), Plant(0.24), Floor(0.19), Chair(0.14), Picture(0.12)
Heads	Picture(0.31), Table(0.24), Wall(0.17), TV(0.16), Floor(0.10)	Table(0.34), Wall(0.23), Picture(0.22), TV(0.14), Cabinet(0.06)
Office	Wall(0.25), Picture(0.24), Floor(0.16), Table(0.11), Chair(0.08)	Wall(0.38), Table(0.14), Shelving(0.12), Picture(0.09), Chair(0.08)
Pumpkin	Cabinet(0.23), Picture(0.23), Wall(0.21), Floor(0.11), Ceiling(0.06)	Wall(0.30), Cabinet(0.23), Ceiling(0.14), Door(0.08), Picture(0.06)
Red Kitchen	Cabinet(0.22), Picture(0.20), Floor(0.19), Wall(0.12), Counter(0.09)	Wall(0.30), Cabinet(0.20), Picture(0.12), Table(0.10), Counter(0.09)
Stairs	Stairs(0.58), Wall(0.21), Picture(0.16), Floor(0.03), Railing(0.01)	Stairs(0.57), Wall(0.34), Floor(0.03), Railing(0.02), Towel(0.02)

ing LSeg features with depth-based fusion. Since VLMaps relies on RGB-D input, comparative experiments are limited to indoor scenes, whereas the proposed approach remains applicable to both indoor and outdoor environments using only monocular RGB inputs.

For both SC-VLMaps and VLMaps, ground-truth camera poses provided by the datasets are used during the mapping stage. We do not estimate odometry in this work. The SCR model is trained only on the training split, and during mapping, the network is frozen and applied to unseen test sequences. Ground-truth poses are used solely for coordinate fusion.

B. Map Qualitative Comparison

Indoor Scene. Fig. 3 presents a qualitative comparison of the Pumpkin scene from the 7Scenes dataset. Fig. 3 (a) shows the ground-truth mesh, while (b) shows the output of VLMaps, and (c) shows the maps generated by our method, SC-VLMaps. In both (b) and (c), the right-hand view illustrates the attention heatmap obtained when querying with the prompt "sofa."

Reconstructions produced by VLMaps often exhibit fragmented geometry and noisy surfaces, particularly along walls and other large planar structures. These distortions arise from the use of RGB-D input, where missing or inaccurate depth measurements pass into the 3D fusion process and reduce the quality of the reconstructed surfaces. Consequently, structural features such as sofas and tables are blurred, and semantic activations appear scattered and poorly localized. Nevertheless, the prompt-based attention mechanism in VLMaps remains functional, as evidenced by its partial activation over the target object.

In contrast, the proposed method, SC-VLMaps, yields denser and structurally coherent reconstructions. By directly regressing scene coordinates from RGB images via a learned model, the system is able to infer geometry in regions where depth is unavailable or unreliable (e.g., reflective or transparent surfaces). These coordinates, when fused across frames, yield maps with sharper object boundaries and enhanced continuity in planar regions. The heatmap corresponding to the "sofa" prompt further demonstrates improved semantic alignment and localization accuracy.

We further provide qualitative comparisons on additional scenes from the 7Scenes dataset in Fig. 4. Across Office, Fire, and Red Kitchen, SC-VLMaps consistently produces reconstructions that are structurally more aligned to the ground-truth geometry compared to VLMaps. In particular, object queries such as "table" and "plant" yield sharper and more localized activations with SC-VLMaps, whereas VLMaps often produces scattered or diffuse responses. It is also worth noting that VLMaps exploits full-depth supervision from the dataset, whereas our method does not rely on depth at all. Instead, the scene coordinate bank is trained only on the training split, and maps are generated online using the test sequences. This separation highlights the generalization capability of SC-VLMaps and explains its superior performance in both reconstruction fidelity and semantic recall.

TABLE IV

COMPARISON OF MAP STATISTICS ON THE OUTDOOR DATASET, CAMBRIDGE LANDMARKS. SINCE VLMAPS RELIES ON DEPTH INPUT, RESULTS ARE REPORTED ONLY FOR OUR METHOD (SC-VLMAPS). HIGHER DENSITY (↑) REFLECTS MORE COMPLETE RECONSTRUCTIONS, WHILE OCCUPIED VOXELS (∼) AND SCENE VOLUME (M^3) ARE PROVIDED FOR CONTEXT.

	Kings College	Old Hospital	Shop Facade	St. Mary's Church
Density (↑)				
SC-VLMaps	0.000182	0.000495	0.000468	0.000967
Occupied Voxels (∼)				
SC-VLMaps	16,321	18,655	35,017	112,111
Scene Volume [m^3]				
SC-VLMaps	429,832.38	53,684.43	106,533.97	165,054.92

Outdoor Scene. To evaluate generalization to outdoor environments, qualitative results are reported on the St. Mary's Church scene from the Cambridge Landmarks dataset (see Fig.5). Fig. 5 (a) shows the ground-truth reconstruction, while (b) shows the reconstructed point cloud generated by SC-VLMaps, and (c) displays the corresponding heatmap for the prompt "window."

Although the reconstructed map is not entirely complete, the architectural structure of the church is clearly recovered, and the semantic activation for the "window" region is well-localized, despite the absence of depth supervision. These results highlight the method's ability to scale to large outdoor scenes and demonstrate its applicability beyond indoor environments.

C. Quantitative Comparison

While qualitative results highlight structural and semantic improvements, we also introduce proxy quantitative metrics to compare map quality. Since our method does not rely on depth sensors, metrics such as navigation success rates (used in VLMaps) are not directly applicable. Instead, we evaluate two aspects: geometric density and semantic alignment.

Our metrics to compare our generated map are as follows:

- Map density. Density is defined as the ratio of occupied voxels to scene volume. Higher density (↑) indicates more complete geometric coverage, while lower occupied voxel counts reflect more compact and less noisy maps.
- Semantic Alignment Score (SAS). SAS measures how well voxel-level embeddings align with a set of language prompts. Given voxel features f_i and prompt embeddings e_j, we compute the cosine similarity matrix $s_{ij} = f_i^\top e_j$. For each voxel, the maximum similarity across prompts is calculated, and SAS is the average of these maximum values. Since the evaluated datasets do not provide dense open-vocabulary semantic ground-truth labels, SAS is not intended to measure semantic classification accuracy. Instead, it evaluates the strength of voxel–language alignment in the shared embedding space, serving as a proxy for open-vocabulary grounding. A higher SAS (↑) indicates stronger semantic alignment.
- Coverage. Coverage@T reports the fraction of voxels whose maximum similarity exceeds a threshold T. Intuitively, it measures the proportion of voxels that carry strong semantic alignment. We evaluate Coverage at 70, 80, and 90. Higher values (↑) are better, reflecting a stronger and more consistent language grounding.

Table I reports the voxel occupancy statistics between SC-VLMaps and VLMaps across scenes in the 7Scenes dataset. From Table I, it shows that SC-VLMaps consistently achieves higher voxel density (↑) in every scene, indicating more complete spatial coverage and fewer gaps in the reconstructed volume. Although VLMaps often records larger raw counts of occupied voxels, these voxels are distributed less efficiently, leading to lower overall density. This also suggests that VLMaps tends to scatter points across the bounding volume, while our method produces more coherent and compact reconstructions. Scene volume remains identical across methods since it is determined by the dataset's bounding box and is not influenced by the mapping algorithm. Taken together, these results demonstrate that SC-VLMaps produces denser and more structurally consistent maps, while avoiding the noisy over-occupancy characteristic of VLMaps.

Subsequently, in Table II, we report the results for SAS and Coverage in all scenes of the 7Scenes dataset. Across all scenes, our method achieves higher SAS than VLMaps, reflecting more reliable voxel–text matching. Coverage further reveals robustness at stricter thresholds: while both methods saturate at Cov@70, VLMaps quickly degrades at Cov@80 and Cov@90, dropping below 0.4 in many cases. In contrast, our method sustains significantly higher coverage, e.g., in the Stairs scene (Cov@90: 0.743 vs. 0.591), demonstrating stronger semantic grounding even in challenging environments.

TABLE V

SEMANTIC ALIGNMENT RESULTS ON THE CAMBRIDGE LANDMARKS (OUTDOOR). VLMAPS CANNOT BE EVALUATED DUE TO ITS RELIANCE ON DEPTH INPUT. REPORTED METRICS INCLUDE THE SEMANTIC ALIGNMENT SCORE (SAS) AND COVERAGE AT THRESHOLDS OF 70, 80, AND 90, WHERE HIGHER VALUES INDICATE BETTER PERFORMANCE.

Scene	SAS	Cov@70	Cov@80	Cov@90
Kings College	0.909	0.995	0.979	0.646
Old Hospital	0.912	0.997	0.961	0.688
Shop Facade	0.889	0.995	0.956	0.456
St. Mary's Church	0.896	0.998	0.954	0.465

We further provide a breakdown of dominant voxel categories as shown in Table III. Table III shows that VLMaps tends to over-emphasize structural classes such as wall and stairs, while our method distributes assignments more evenly across semantically meaningful objects such as tables, pictures, and cabinets. This diversity indicates that our maps preserve richer semantic cues, enabling more fine-grained text queries.

We also provide the voxel statistics of the map, SAS, and Coverage value over the Cambridge Landmarks dataset in Table IV and Table V. On the outdoor Cambridge dataset, only our method (SC-VLMaps) can be evaluated, as VLMaps relies on RGB-D input. The voxel statistics (Table IV) show that our reconstructions achieve non-trivial densities across all four scenes, with tens of thousands of occupied voxels distributed over large spatial volumes. Table V further indicates strong consistency between regressed coordinates and language features: average SAS scores remain above 0.89 across scenes, with high coverage even at stricter thresholds (e.g., Cov@90 above 0.45). These results highlight that SC-VLMaps scales beyond constrained indoor settings and produces coherent, semantically aligned reconstructions even in challenging outdoor environments.

Overall, these results show that bypassing depth sensing not only yields cleaner geometry but also strengthens semantic alignment. This advantage stems from regressing scene coordinates directly from RGB, which allows the system to infer structure even where depth is missing or unreliable, and from training the scene coordinate bank only on the training split, ensuring generalization to unseen test sequences. In addition, because each fused voxel combines both geometric position and semantic descriptors, the resulting maps carry stronger voxel–language alignment. Together, these factors enable SC-VLMaps to avoid the sparsity and noise inherent in depth-based fusion, producing maps that are more compact, semantically diverse, and robust to strict alignment thresholds—making them better suited for text-driven interaction in both indoor and outdoor domains.

D. Limitations and Future Work

SC-VLMaps demonstrates strong performance on both indoor (7Scenes) and outdoor (Cambridge Landmarks) benchmarks, but several limitations remain. Our evaluation focuses on voxel density and semantic alignment as interpretable proxies for map quality, which do not fully capture downstream utility in tasks such as navigation or interaction. Although SC-VLMaps is built on scene coordinate regression (SCR), which naturally supports camera relocalization via predicted 2D–3D correspondences and a PnP-based solver, relocalization performance is not explicitly evaluated in this work.

While the method generalizes to outdoor environments, extremely large or open scenes may still produce incomplete reconstructions due to sparse viewpoints or textureless regions. In addition, the framework assumes accurate camera poses and relies on frozen language features (LSeg), which may limit robustness in unconstrained real-world settings.

Future work will extend SC-VLMaps in three directions: (i) joint learning of pose estimation and mapping for improved robustness, (ii) evaluation in interactive tasks such as robotic navigation and semantic retrieval, and (iii) hierarchical or adaptive voxel structures to better handle large-scale outdoor scenes.

V. CONCLUSION

We presented SC-VLMaps, a depth-free visual–language mapping framework that extends multi-scene learned scene coordinate regression with semantic embeddings from frozen language models. Unlike prior methods such as VLMaps, which rely on RGB-D input and are limited to indoor environments, SC-VLMaps operates directly from monocular RGB images and generalizes to both indoor and outdoor scenes.

Experiments on 7Scenes and Cambridge Landmarks demonstrate that our approach produces denser and more coherent maps, with stronger semantic alignment to language prompts. Quantitative results highlight consistent improvements in voxel density and Semantic Alignment Score, while qualitative comparisons show sharper structures and more localized semantic activations.

By bypassing explicit depth sensing and leveraging reprojection-based supervision, SC-VLMaps achieves compact, semantically diverse, and structurally consistent maps suitable for text-driven queries. This work extends language-conditioned mapping to more unconstrained environments and provides a foundation for future work in robust pose estimation, task-driven evaluation, and large-scale outdoor mapping.

REFERENCES

[1] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[2] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=RriDjddCLN

[3] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3674–3683.

[4] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *2017 IEEE International Conference on Robotics and automation (ICRA)*. IEEE, 2017, pp. 4628–4635.

[5] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 740–10 749.

[6] M. Runz, M. Buffier, and L. Agapito, "Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects," in *2018 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 2018, pp. 10–20.

[7] G. Narita, T. Seno, T. Ishikawa, and Y. Kaji, "Panopticfusion: Online volumetric semantic mapping at the level of stuff and things," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4205–4212.

[8] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3D scene graph construction and optimization," 2022.

[9] N. Hughes, Y. Chang, S. Hu, R. Talak, R. Abdulhai, J. Strader, and L. Carlone, "Foundations of spatial perception for robotics: Hierarchical representations and real-time systems," *The International Journal of Robotics Research*, 2024. [Online]. Available: https://doi.org/10.1177/02783649241229725

[10] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Advances in neural information processing systems*, vol. 26, 2013.

[11] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[13] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7086–7096.

[14] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2930–2937.

[15] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac-differentiable ransac for camera localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6684–6692.

[16] E. Brachmann, T. Cavallari, and V. A. 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, pp. 5044–5053.

[17] S. T. Nguyen, A. Fontan, M. Milford, and T. Fischer, "Focustune: Tuning visual localization through focus-guided sampling," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3606–3615.

[18] F. Wang, X. Jiang, S. Galliani, C. Vogel, and M. Pollefeys, "Glace: Global local accelerated coordinate encoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 562–21 571.

[19] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R2former: Unified retrieval and reranking transformer for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 370–19 380.

[20] E. Brachmann and C. Rother, "Visual camera re-localization from rgb and rgb-d images using dsac," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5847–5865, 2021.

[21] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.