UNIVERSITY OF
LATVIA

Summary of
Doctoral Thesis

Peteris Racinskis

PERCEPTION AND CONTROL IN AUTONOMOUS ROBOTICS USING VISION-LANGUAGE EMBEDDINGS

Riga 2026

UNIVERSITY OF
LATVIA

FACULTY OF SCIENCE AND TECHNOLOGY

Peteris Racinskis

PERCEPTION AND CONTROL IN AUTONOMOUS ROBOTICS USING VISION-LANGUAGE EMBEDDINGS

SUMMARY OF DOCTORAL THESIS

Submitted for the degree of Doctor of Science (Ph.D.)
in Computer Science
Subfield of Intelligent Systems

Riga 2026

The Doctoral Thesis was developed at the Robotics and Machine Perception Laboratory, Institute of Electronics and Computer Science (EDI) from 2023 to 2025.

The Thesis contains an introduction, 4 chapters, a summary, key results and conclusions, a bibliography, supplemental information, and acknowledgments.

Form of the Thesis: a dissertation.

Scientific supervisor: *Dr. sc. comp.* **Modris Greitans**, senior researcher, Institute of Electronics and Computer Science (EDI), Latvia.

Reviewers:

1. Professor, *Dr. sc. comp.* **Inguna Skadiņa**, University of Latvia, Latvia;
2. Professor, PhD **Vytautas Bučinskas**, Vilnius Gediminas technical university, Lithuania.
3. Professor, PhD **Po Ting Lin**, National Taiwan University of Science and Technology, Taiwan.

Defense of the Doctoral Thesis will take place at a public session of the Doctoral Committee of Computer Science, Informatics, Electrotechnics, Electronics and Communication Technologies of the University of Latvia on January 23, 2026.

The Doctoral Thesis and its Summary are available at the Library of the University of Latvia, Kalpaka Blvd 4, Riga.

Chairman of the Doctoral Committee:	Dr. sc. comp. Guntis Bārzdiņš
Secretary of the Doctoral Committee:	Sintija Siliņa

ISBN 978-9934-36-481-5
ISBN 978-9934-36-482-2 (PDF)

ABSTRACT

This doctoral thesis presents advancements in perception and control for autonomous robotics, integrating vision-language semantics to address challenges in unstructured environments. The research focuses on developing semantic Simultaneous Localization and Mapping (SLAM) systems. Key contributions include the application of visual open-set semantics within large-scale LiDAR maps for terrain segmentation in autonomous robot navigation, and the integration of Large Language Model (LLM)-based high-level planning with these semantic maps in a tabletop object stacking scenario. The study aims to fuse various sensor data — LiDAR, depth camera, color images, and GNSS — into a unified, language-grounded environmental model, facilitating human-robot interaction through voice or text. Experimental results also validate novel data set collection methodologies, demonstrating that visual fiducial markers positioned using surveying tools can substitute for GNSS-INS-based ground truth position and orientation estimates in scenarios when these are unavailable or inadmissible. The results of this work have been published across three first-author articles in indexed journals, two international conferences and a project report technology description publication. This work introduces two publicly available resources — the EDI-SLAM dataset and the open-source SLAMVDB software package. Two demonstrators developed in this dissertation — a semantic perception system and static manipulation-focused robot control architecture — have been positively assessed by industry representatives.

Keywords: Robotics, Computer Vision, SLAM, Perception, LLM, VLM

CONTENTS

LIST OF ABBREVIATIONS

EDI - Institute of Electronics and Computer Science, Latvia

SLAM - Simultaneous Localization and Mapping

LiDAR - Light Detection and Ranging (laser range scan)

VLM - Vision-Language (embedding) Model

LLM - Large Language Model

NLP - Natural Language Processing

GNSS - Global Navigation Satellite System

ECEF - The ”Earth Centered, Earth Fixed” global Cartesian coordinate system

ROS, **ROS2** - Robot Operating System, Robot Operating System 2

RGB - ”Red, Green, Blue”, used for both the sensor type and data modality

RGB-D - ”Red, Green, Blue, Depth”, used for both the sensor type and data modality

RTK - Real-Time Kinematic GNSS

ATE - Absolute Trajectory Error

RPE - Relative Pose Error

ICP - the Iterative Closest Point algorithm

HLP - High-level Planning subsystem

LLP - Low-level Planning subsystem

SotA, **BSotA** - State of the Art, Beyond State of the Art

PnP - the Perspective-n-Point problem or an algorithm solving it

INTRODUCTION

Across foundational industries such as manufacturing, agrifood, and forestry, a confluence of demographic shifts, evolving job requirements, and economic pressures is fueling concerns over a looming labor shortage and a deepening skills crisis. Recent surveys and industry reports consistently highlight these challenges, indicating that businesses globally are struggling to find and retain workers, particularly those possessing the specialized or emerging skills needed for modern operations [1, 2]. Historically, robotics has long been seen as a potential solution to many types of shortages in labor markets. In repetitive, static tasks, such as the ones faced by large-scale manufacturing enterprises, this has in fact been borne out through static industrial manipulators. However, though ubiquitous, these are still largely programmed by hand and execute pre-programmed trajectories [3], which limits their flexibility and therefore potential uses. Production lines are still being set up by highly skilled and scarce technicians, and there is little natural feedback or interactivity between a robot system and the typical line worker. Leaving behind the confines of the production line, the greatest strides towards autonomy have been made in relatively structured environments, such as the road conditions faced by autonomous vehicles [4]. As application domains trend towards the less structured, one can observe a decrease in technology readiness levels (TRLs) for perception and control technologies through semi-structured (e.g. agricultural [5]) and nearly unstructured (such as forestry [6]) environments. One can identify some key challenges that explain this pattern:

1. Modeling environments for perception and planning becomes increasingly difficult as fewer abstractions can be assumed ahead of time;

2. Setting up a robot control system becomes increasingly expensive and time consuming as the nature of the task grows less repetitive.

Addressing both of these challenges is required to open up a path towards deploying robots to conduct open-ended tasks in difficult environments. Potential applications range from the encouraging the use of manipulator arms in low volume manufacturing or the service industry to autonomous employment of mobile robots in use cases featuring complex, unstructured environments, such as agrifood, forestry, infrastructure maintenance and defense.

Scientific novelty

This dissertation covers a corpus of research and engineering work largely focused on development of semantic SLAM systems, using them in industrial robot control and the development of hardware, software and methodology sur-

rounding these goals — from data collection to open-source software for public release. The primary scientifically novel contributions, however, are:

- Use of visual open-set semantics in a large-scale LiDAR map for terrain segmentation, intended for uses in ground-based autonomous robot navigation;
- The integration between LLM-based high-level planning techniques and open-set semantic maps;
- Methodology for testing and evaluating SLAM systems in GNSS-denied environments using visual fiducial markers and surveying equipment.

Aim and objectives

The general motivation and **aim** of the study is advancing autonomy and human interactivity in both stationary and mobile robots through the fusion of various sensor observations — LiDAR or depth camera range observations, color images, GNSS — into a unified, language-grounded model of a robot's surroundings. The results of this study should enable future autonomous robot control systems to better navigate complex, unstructured environments such as forests, and allow human operators to more intuitively interact with robots.

The specific research **objectives** of the study are to:

1. Develop an open-set semantic mapping system robust to operation in unstructured outdoor environments and capable of operating over long trajectories and/or time horizons;
2. Demonstrate the capability of the mapping system to perform various lookup operations relevant in the industrial manipulation and outdoor rover context — object pose estimation, terrain segmentation;
3. Create the tooling and methodology necessary to test, verify and validate system performance under relevant conditions.

Main theses

To structure and motivate the numerous contributions made during the development of this dissertation, the following three theses are put forward, which will be supported by the results described thereafter:

- **Thesis 1** — visual fiducial markers whose corner points are located using landscape surveying equipment provide a GNSS-independent reference pose measurement method that outperforms RTK-corrected multi-band GNSS in environments with signal degradation.

- **Thesis 2** — open-set semantic voxel or depth maps achieve object recall rates in excess of 90% in a centimeter resolution tabletop setting, and an LLM-based command parsing pipeline integrating such maps successfully understands and executes simple stacking instructions.

- **Thesis 3** — an open-set voxel map containing projected vision-language model image embedding features achieves voxel-wise terrain classification accuracy as high as 87% when a heuristic to dismiss void observations is implemented.

Short description of the methods

To motivate and guide the subsequent research, the Author has conducted a literature review in the field of perception, environment modeling, localization and map construction for autonomous robotics [7]. The Author developed a hand-portable sensor package for collecting synchronized LiDAR, IMU, GNSS and image data, alongside the software necessary to run it and tooling for processing the related data. A key contribution was the development of the visual fiducial marker annotation method described in [8], and this work has culminated in the collection of the *EDI-SLAM* data set [9]. To test the viability of open-set semantic mapping as a method for robot perception, an indoor industrial robot demonstrator was developed for conducting tabletop pick-and-place tasks [10]. Tooling for manual annotation of tabletop scenes and semi-automatic creation of ground-truth action plans for testing the HLP performance were developed and used to quantitatively evaluate system performance. Finally, a second demonstrator in the form of an open-source software package was developed by the Author for conducting Open-set Semantic SLAM in outdoor environments and building large-scale voxel grid maps with interfaces for terrain segmentation and object search. A version of this has been publicly released as *SLAMVDB* [11]. Since much of the work herein was conducted under the Latvian State Research program VPP-EM-FOTONIKA-2022/1-0001, a technical project report document containing implementation details for both demonstrators can be found at [12].

Publication and presentation of results

Publications and conference abstracts relevant to this Doctoral thesis:

(I) Racinskis, Peteris, Janis Arents, and Modris Greitans. ”Constructing maps for autonomous robotics: An introductory conceptual overview.” Electronics 12, no. 13 (2023): 2925.

(II) Racinskis, P., Arents, J., Greitans, M. Towards a Multi-modal, Multi-layer Mapping Framework for Autonomous Robotics–an Outline, International Workshop on Embedded Digital Intelligence (IWoEDI’2023), Riga, Latvia. **(two-page abstract)**

(III) Racinskis, Peteris, Janis Arents, and Modris Greitans. ”Annotating SLAM data sets with Apriltag markers.” In 2024 10th International Conference on Automation, Robotics and Applications (ICARA), pp. 438-442. IEEE, 2024. **(five-page report)**

(IV) Racinskis, Peteris, Oskars Vismanis, Toms Eduards Zinars, Janis Arents, and Modris Greitans. ”Towards Open-Set NLP-Based Multi-Level Planning for Robotic Tasks.” Applied Sciences 14, no. 22 (2024): 10717.

(V) Racinskis, Peteris, Gustavs Krasnikovs, Janis Arents, and Modris Greitans. ”The EDI Multi-Modal Simultaneous Localization and Mapping Dataset (EDI-SLAM).” Data 10, no. 1 (2025): 5.

The results detailed in this dissertation have been presented at the following international scientific conferences and research-industry forums:

(a) Two-page abstract: 2023 International Workshop on Embedded Digital Intelligence (IWoEDI’2023). Riga, Latvia, 2023;

(b) Conference paper: 2024 IEEE 10th International Conference on Automation, Robotics and Applications (ICARA). Athens, Greece, 2024;

(c) Poster: 2024 IEEE 14th International Conference Nanomaterials: Applications & Properties (NAP). Riga, Latvia, 2024;

(d) Poster and physical demonstrator: *5g Techritory*. Riga, Latvia, 2024;

(e) One-page abstract, poster, physical demonstrator: European Robotics Forum 2025. Stuttgart, Germany, 2025.

The research in this dissertation was primarily funded through two project grants:

- Latvian state research program No. VPP-EM-FOTONIKA-2022/1-0001 “Viedo materiālu, fotonikas, tehnoloģiju un inženierijas ekosistēma”
- EdgeAI “Edge AI Technologies for Optimised Performance Embedded Processing” project, which has received funding from CHIPS JU under grant agreement No 101097300.

The Author’s contribution

In all publications the Author was the primary author, and prepared most or all of the manuscript. In (I), (II), (III) the Author conducted all of the research and development activity. In publication (IV), the Author conceptualized the idea, developed the system architecture, oversaw the development, deployment

and experimental evaluation of the demonstrator, and implemented the following subsystems directly: open-set semantic map, image data middleware, experimental data collection, data set annotation tooling, evaluation statistics computation. Colleague Oskars Vismanis developed the low-level robot action primitive subsystem. Colleague T.E. Zinars developed the high-level planning subsystem. Both colleagues aided in manuscript preparation by writing the initial draft versions of the sections corresponding to their subsystems. Descriptions of the components developed by them have been used with permission, where relevant. In (V), the Author conducted all of the research and development activity, but colleague Gustavs Krasnikovs recorded most of the sensor tracks that were included in the actual public release of the data set, as well as performing sensor package calibration, set-up and maintenance tasks. In all publications, senior researcher, EDI robotics research group lead Ph.D. Janis Arents and senior researcher, EDI director, doctoral thesis supervisor Dr.sc.comp. Modris Greitans assisted in an advisory and administrative capacity.

Thesis outline

Due to the fact that a considerable amount of material has already been published by the author regarding the research described in this dissertation, much of the content of this document is given in the form of an abbreviated summary, augmented by the addition of some more recent references, results and developments. This thesis contains 30 figures, 10 tables, and 2 algorithm descriptions.

Chapter 1 — Background is a review of the key concepts, research and current state of the art in the fields covered by this disseration. It is largely a combined summary of the background sections of articles (I), (IV), (V) and the project report at [12], with some more recent results included where relevant to bring the information up to date with the latest developments in the field;

Chapter 2 — Sensor Package and Dataset Collection describes the sensor package developed to collect data outdoors, the novel reference pose measurement method, experimental assessments of its accuracy and a comparison with the on-board GNSS-INS track, as well as the *EDI-SLAM* data set;

Chapter 3 — Tabletop Manipulation Demonstrator goes into technical detail regarding the robot control system architecture developed to take advantage of open-set semantics in a static environment, the evaluation of the LLM-based instruction parsing and planning system, as well as results in end-to-end execution;

Chapter 4 — Semantic Perception Systems details the sequence of open-set semantic mapping systems developed as part of this research, their design aspects and the various features added in each iteration, as well as verification methods and results in three domains — object recall, localization accuracy and terrain segmentation.

Summary of the key results

A portable sensor package was built and a large, public data set - *EDI-SLAM* - containing several kilometers of recorded trajectories from hand-carried and vehicle-mounted vantage points was collected. The accuracy of the novel ground truth pose estimation method was employed, and its accuracy experimentally assessed — with a maximum position error under $5cm$ and maximum orientation error under 0.5° when using 4 markers per gate. In parallel, an indoor multi-level robot control system using open-set semantics for object grasp pose estimation was implemented. The system proved capable of recovering over 90% of simple object stacking action plans, and executing two thirds of commands with complete success on an industrial robot arm. A sequence of open-set semantic perception systems was developed, demonstrating object detection and grasp pose estimation for manipulation (up to 94.69% recall, using an octree map and ternary quantization of semantic features) as well as terrain segmentation for navigation (up to 87.35% voxel classification accuracy on the *RELLIS-3D* data set, track *00000*).

1. BACKGROUND

For an autonomous robot to perform meaningful tasks in the world, it must be able to perceive its surroundings, understand its own position and orientation, and execute planned actions. In most currently existing frameworks, low-level controllers actuate individual joints or wheels, and high-level planners generate sequences of actions to achieve a goal [13, 14]. However, these still require precise, programmed setpoints — either in terms of actuator state, or precisely defined trajectory in joint or Cartesian space. To generate these movement targets dynamically, a crucial link between planning and action is the robot's perception system, which provides the environmental model necessary for tasks like navigation and obstacle avoidance [15], while more complex behavior and natural language interfaces require higher-level control.

1.1. Perception and Mapping Systems

When a robot operates without a prior model of its environment, it must solve the **Simultaneous Localization and Mapping (SLAM)** problem — concurrently building a map of its surroundings while tracking its own pose (position and orientation) within that map [16, 17]. This is formally a Bayesian inference problem, aiming to estimate the robot's trajectory and the map structure given a sequence of sensor measurements [16]. Two dominant paradigms exist for solving the SLAM problem. Filter-based approaches, like the Extended Kalman Filter (EKF), estimate only the most recent robot state, marginalizing out past information to maintain computational tractability [18–20]. While effective for real-time tracking or odometry, this marginalization can lead to drift over long trajectories [16]. In contrast, smoothing approaches retain the entire history of robot poses and sensor measurements, optimizing them jointly to find a globally consistent solution [17], and potentially including loop closures — constraints added when the robot recognizes a previously visited location [21–23]. Smoothing SLAM problems are often represented as factor graphs, where robot poses and map landmarks are variables, and with sensor observation constraints (factors) between them [24]. Solving the graph corresponds to a non-linear least squares optimization problem [17].

The output of a SLAM system is typically a *metric* map, such as a point cloud or an occupancy grid [25], which encodes the geometry of the environment. While essential for localization and basic navigation, purely metric maps lack the higher-level understanding required for complex interaction and planning [26]. Decades of research have therefore focused on augmenting maps with semantic layers representing the meaning of objects and areas [27, 28]. The semantic information for these maps is overwhelmingly sourced from computer vision models processing camera imagery. Tasks like object detection and pixel-wise **semantic**

segmentation assign discrete class labels (e.g., ”chair,” ”table”) to parts of an image [29]. A recent and powerful evolution of this is **open-set semantics**, which moves beyond a fixed set of classes. By leveraging joint text-image embedding spaces from Vision-Language Models (VLMs) like CLIP [30], these systems can identify and query for objects and concepts described in natural language, without needing to be retrained for new categories [31, 32]. This capability is crucial for creating flexible and human-interpretable robotic systems.

A particularly challenging application of semantic mapping is **terrain segmentation** for navigation in unstructured, outdoor environments. Unlike typical object recognition, this task requires distinguishing between traversable and non-traversable surfaces (e.g., “*grass*,” “*gravel*,” “*mud*”). Existing datasets for this domain are significantly smaller than those for common objects [33, 34], and model architectures that excel at object segmentation do not always generalize well to the nuanced, large-scale awareness needed for terrain analysis [35]. This thesis posits that by fusing local, open-set semantic observations into a global 3D map, a generic vision model can be effectively adapted to this specialized task.

1.2. High-Level Planning and Control

The ultimate goal of a perception system is to inform action. In modern robotics, there is a strong research trend towards using **Large Language Models (LLMs)** for high-level task planning [36, 37]. These systems can parse natural language commands and generate a sequence of steps for the robot to execute. Some approaches use an LLM to generate a complete plan upfront, which is then validated and executed [38, 39], while others use the LLM iteratively to select the next best action from a predefined set of skills based on the current context [40]. A significant limitation of many current LLM-based planners is that their ”world model” is often confined to the robot’s immediate sensory input or a short history of observations [40, 41]. They largely fail to leverage the rich, persistent, and globally consistent world representation that a SLAM system can provide. This creates a gap: on one hand, we have sophisticated SLAM systems building detailed semantic maps; on the other, we have intelligent planners that cannot fully exploit them. Bridging this gap by developing a large-scale, queryable semantic map designed to interface with modern planning systems is a primary motivator for the work presented in this thesis.

1.3. Challenges, Goals & BSotA

To motivate the research work described in this Doctoral thesis, it is important to summarize key insights and identify some important problems that remain to be addressed, as well as the aspects of capability and performance where the current State of the Art (SotA) is lacking. Practical field trials such as [42], where a number of systems [18, 19, 21, 22, 43] were evaluated, show that large-scale robust mapping of outdoor environments remains a problem. It is therefore neces-

sary to ensure that a robust, reliable and module localization system be deployed at the core of any perception stack. To ensure this, the ability to collect new sensor data to adequately reflect the studied target applications is of critical importance, to avoid overfitting a limited set of available benchmarks. Less often represented environment types — partially or wholly unstructured —- and long trajectories (hundreds of meters, kilometers) need to be collected and tested against. When moving between open spaces with clear, unobstructed GNSS signals and indoor or highly built-up areas, a ground truth annotation method that is independent from the GNSS signal should also be provided. With a GNSS-independent reference pose measurement, it is possible to fairly assess the localization performance of SLAM frameworks that directly fuse GNSS data into their estimates. Bearing in mind how SotA approaches to LLM or multi-modal model-based planning either process only the currently visible scene [40] or a sequence of images the robot has seen [41], but fail to take advantage of SLAM technologies, integration between LLM-based high-level planning and structured semantic environment models must be explored. Finally, in terrain segmentation, where previous approaches focus primarily on improved domain-specific semantic segmentation in images [35] or point clouds [44], this thesis explores the possibility of re-using the same general-purpose open-set image segmentation model in finding traversable surface paths for robot navigation.

Taken all together, the author has thus concluded that the implementation of several technology blocks and experimental activity campaigns was required:

- Perception System — the development of an open-set semantic mapping system, for use in robot perception. An approach had to be identified that is capable of both object detection and pose estimation for manipulation tasks in a confined workspace (thesis 2), as well as terrain segmentation on a large scale, to enable navigation in autonomous, mobile robots (thesis 3);

- Control Integrations — to validate the applicability of the aforementioned mapping system in actual robotic control problems (thesis 2), an architecture for the broader system that can take advantage of such a map needed to be developed, the requisite modules for high- and low-level planning and control had to be built, interfaces between these had to be implemented and the capabilities of the resulting system had to be validated in real-world robot control;

- Data Provision — to ensure that appropriate testing, verification and validation data are available, methodologies and instrumentation for data set collection, ground-truth labeling and employment in quantitative evaluation needed to be developed (thesis 1).

2. DATA SET AND REFERENCE POSES

In order to provide the data that are used in developing, testing and evaluating the perception systems in this dissertation, a sensor set-up and collection methodology were needed. At the outset of the research activity described in this dissertation, EDI's robotics laboratory already possessed a two-robot workstation equipped with multiple depth cameras. However, no equipment was available for collecting LiDAR or GNSS data along arbitrary outdoor trajectories through unstructured environments. Therefore, a portable sensor package was developed. In addition, to allow localization accuracy benchmarking in GNSS-denied environments, or for systems that use GNSS, an independent reference pose measurement method was introduced and its accuracy was assessed w.r.t. static motion capture equipment. The sensor package and new reference pose measurement method were then employed in collecting the *EDI-SLAM* [9] data set. This was used in a localization error measurement for the on-board GNSS-INS sensor, as well as evaluating the performance of some of the SLAM systems used in the semantic mapping technologies.

2.1. Sensor Package

Taking notes from the sensor package used in recording the *TUM-VI* [45] visual-inertial odometry benchmark, a hand-portable, adjustable mounting frame was constructed to hold the sensors, compute module, GNSS antenna and power delivery circuitry, shown in Figure 2.1. LiDAR was selected over a depth camera for its higher range and tolerance of outdoor lighting conditions. Cameras were mounted primarily to enable the use of computer vision in semantic inference; two are required to also allow the testing of stereo and stereo-inertial SLAM systems, in addition to future proofing for more reliable stereo-to-depth estimation methods — which may eventually allow for doing away with the LiDAR in some workflows. A GNSS-INS navigation system is to be mounted to serve as the primary dense (high-frequency and high-availability) ground truth annotation method in scenarios where GNSS drift is expected to be low, such as open field trajectories.

The sensor package mounts an *Ouster OS1 rev7* [46] 32-line mechanical LiDAR (1), which outputs point clouds at 10Hz and IMU data at 100Hz; $2\times$ *Basler Dart 1920-160uc* [47] global shutter USB3.0 cameras (2), equipped with fixed focal length $4mm$ 1/1.8" lenses; an *Xsens MTi-680g* [48] inertial navigation unit (occluded in the image) with a multi-band *u-blox ZED F9 RTK GNSS* receiver(3); and an *Intel nuc* compact PC, running *Ubuntu Linux 20.04 LTS* and ROS1 *Noetic*, which performs data recording (4). Reflective markers (7) can be affixed to the package for testing and calibration w.r.t. the *Optitrack* motion tracking system.The battery is carried separately — in a backpack for hand-held

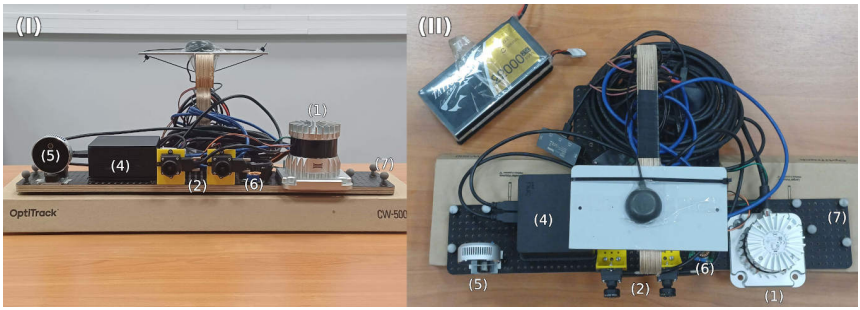

Figure 2.1: (I) front and (II) top views of the sensor package, with some of the sensors indicated. (III) one of the images processed for reference pose measurement in the *courtyard_gt* track, showing a tracking gate with detected marker indices and their corners drawn in black.

recording, in the cabin when recording using a car and fastened separately on the UGV platform, with detachable extension cables used to supply power in the various configurations. A custom *Pylon* camera ROS driver was implemented, to provide a software-synchronized LiDAR-camera capture mode, invoking the software trigger on both cameras whenever a point cloud is published. The *kalibr* software toolkit [49] is used to obtain the various camera intrinsic, camera-to-camera, camera-to-LiDAR and camera-to-IMU extrinsic calibration parameters.

2.2. Reference Pose Measurement

Given the requirement for statistically independent evaluation of SLAM systems that directly fuse the same GNSS data into their state predictions, as well as sky occlusion and multipath scattering and occlusion concerns, an entirely independent method of ground truth pose measurement must also be provided. A prominent method for providing such alternate measurements used in other data sets is the use of motion capture systems, e.g. *Optitrack* system [45, 50]. While they do provide entirely GNSS-independent measurements, and sub-milimeter level accuracy, such installations are extremely limited in their effective working area by the need to deploy numerous (highly expensive) cameras around the body of interest to provide multiple views of its reflective marker pattern.

In this dissertation, the Author introduces an alternative measurement system for SLAM data sets incorporating a camera feed [51]. *Apriltag2* [8] fiducial markers attached to marker plates are placed along the envisioned camera track, then traditional surveying tools — a total station — are used to measure the positions of the marker plates, either in a floating Euclidean frame or, if possible, in a geo-referenced coordinate system (using independently obtained GNSS-RTK measurements from different equipment) such as ECEF. Given known camera

calibration parameters and the external reference frame corner point locations, a camera pose is independently estimated for every frame containing marker observations, using the *perspective-n-point* algorithm [52]. In the typical deployment configuration, the visual markers are arranged in "tracking gates", each of which consists of two backing plates mounting two markers each, to present a total of 16 corners for the PnP pose estimation algorithm.

The primary advantage of this set-up *vis-a-vis Optitrack* and similar systems is that it can be deployed practically anywhere, since the visual tracking gate deployments are much simpler and less expensive than motion capture installations. Furthermore, any pose estimates are aligned in time with the camera frames by construction, without the need to conduct any additional frame-to-pose synchronization. Compared to GNSS-INS, the main advantages are the aforementioned statistical independence as well as the ability to deploy such a system indoors or in other locations where the satellite signal is unavailable or unreliable. The most notable disadvantages are reduced coverage — pose estimates are available only near surveyed tracking gates — and before any recording can be performed, this placement and surveying work needs to be done, increasing the time and cost of collecting trajectories in new locations.

2.3. Data Set

Using this sensor package, the driver software and calibration methodology At the time of writing this document, the data set is in its first major public release, an accompanying publication for which can be found at [9]. As of the writing of this document, the release version contains recordings with and without the reference pose estimates, collected over five runs in three different environments:

- the EDI courtyard, some of its surroundings — a cityscape with some stretches of dense vegetation and shrub mixed in. The high buildings in the courtyard degrade GNSS signal quality through obstructing line of sight and multipath scattering (2 tracks);
- a flat, grassy field — featuring high-quality GNSS data but relatively featureless segments which make complicate scan-to-scan or image-to-image tracking (2 tracks);
- a network of forest service roads — straight, long dirt and gravel tracks under the forest canopy, featuring highly repetitive scenery (which aggravates the risk of spurious loop closures) and intermittent GNSS availability (1 track).

When performing ground-truth-annotated recording runs, *Apriltag* tracking gates were placed in various locations along the planned route. The trajectory always starts with the first gate visible. Others are spaced approximately

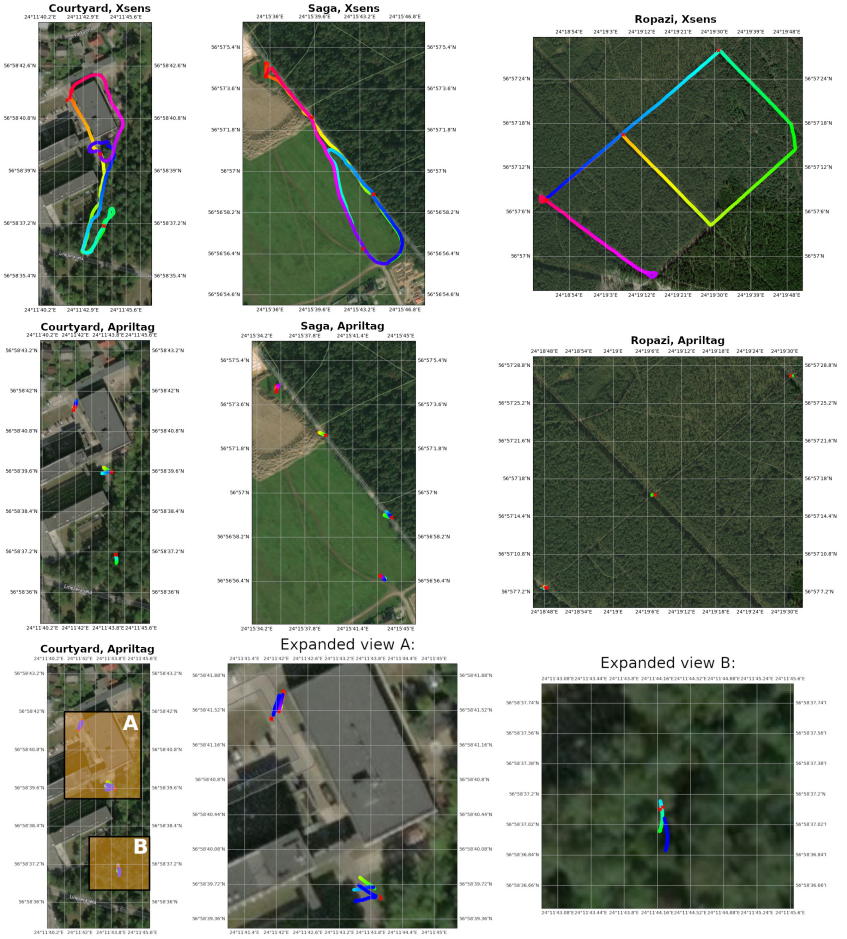

Figure 2.2: Satellite maps with reference positions — made with GNSS-IMU (top) and visual marker tracking gates (middle). Expanded views (bottom) of marker-derived reference poses (solid blue), markers (red) and corresponding GNSS-IMU estimates (gradient)

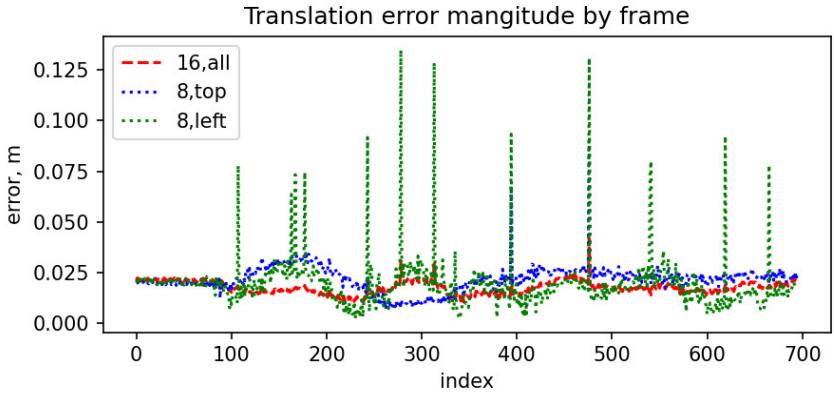

Figure 2.3: Visual marker pose estimate deviation from *Optitrack* w.r.t. observation index number, with different marker configurations.

evenly along the envisioned path, to ensure any significant drift accumulation gets measured. Additionally, with the EDI courtyard track, an effort has been made to place the tracking gates in occluded spots, where it was expected that significant GNSS degradation would be encountered. The sensor package was carried in two different ways during collection — in the courtyard and one of the recordings made in the open field, the package was carried in-hand. For the much longer forest road track, as well as the other field tracks, the rig was mounted to the roof of a car, to allow for longer recordings. 3 of the tracks feature reference pose annotations — one for each environment. Satellite images of the these are shown in Figure 2.2. The data set is hosted under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License [53] on EDI's domain [54].

2.4. Results

To ensure the quality of the ground truth reference pose measurements included in the *EDI-SLAM* data set [9], assessments of both the newly proposed surveyed *Apriltag* localization method and the on-board GNSS-INS trajectory were performed, the results of which are collected in Table 2.1. The former was assessed w.r.t. a fixed *Optitrack* motion capture system. Several marker arrangements were tested and, as visible in Figure 2.3, using more corner points spaced farther apart in image space is important for reducing the incidence of outlier errors. Using 4 marker plates, average angular error of $0.43°$ and translation error of approx. 1.8cm were obtained, up to a bias in the calibration used — with maximum values of approx. $2°$ and 4.2cm, respectively.

An assessment of the GNSS-INS system error has also been made on the

Table 2.1: Reference pose measurement results.

Marker count and arrangement	**Evaluation metric**			
	mean ϵ_θ	max ϵ_θ	mean ϵ_t	max ϵ_t
all 4 markers, 16 points	$0.43°$	$2.04°$	0.018m	0.042m
top 2 markers, 8 points	$0.42°$	$2.39°$	0.022m	0.090m
left 2 markers, 8 points	$0.59°$	$4.51°$	0.019m	0.135m

Track	ϵ_{rmse}	$\epsilon_{rmse,XY}$	ϵ_{ate}	$\epsilon_{ate,XY}$	ϵ_{rpe}
courtyard_gt	77.291	6.873	4.000	3.193	0.008
saga_gt	1.305	0.452	0.114	0.099	0.024
ropazi_gt	8.223	2.165	3.059	2.405	0.047

EDI-SLAM data set, summarized in the bottom section of Table 2.1. These results show that the GNSS-INS estimates largely agree with the marker-based positioning on the open-sky *saga_gt* track — with an RMSE of less than $50cm$ in the XY plane, a metric that quantifies raw error in individual measurements without computing any kind of alignment beforehand — there is a substantial divergence on the *courtyard_gt* track, where the sky is often occluded and the presence of tall buildings introduces difficult to predict reflections or scattering. The largest component of the error is a systematic bias in altitude estimates, but discontinuities are also present, which distort the aligned ATE estimates.

The tradeoffs between the new reference pose measurement method and GNSS-INS are clearly visible in Figure 2.2. The top row shows the much denser coverage afforded by the high-sample-rate and near-continuously present *Xsens* track, where the visual estimates are only available near the tracking gates. However, when examining the *courtyard_gt* track, where the greatest amount of GNSS-INS deviation from *Apriltag* poses has been measured numerically, a clear trajectory deviation is apparent — even though the largest component of the error is along the altitude axis, not visible in the satellite images.

3. TABLETOP STACKING DEMONSTRATOR

The main research goal at the robot control stack level is to create a system that can process a free-form natural language prompt to turn it into a sequence of motion commands to a real robot, achieving the desired effect in the robot's environment. In accordance with the aims and objectives of the dissertation, the decision to pursue an approach integrating an explicit, structured map of the environment was taken early on. To ensure modularity and portability, action primitives — robotic skills — have been selected as the guiding paradigm in control, implemented through a . A high-level planning subsystem, using LLMs at its core, is employed to convert unstructured commands in natural language into action plans the robot can execute. The proposed architecture has been implemented on a real industrial manipulator and tested both component-wise and in end-to-end execution.

3.1. Multi-level Control System

To take full advantage of the capabilities offered by SotA LLMs and multimodal models, while also making use of a structured model of the environment, the following three-block architecture has been devised for planning, control and perception:

- **High-level Planning (HLP)** subsystem — the main NLP and user interaction driver, processing queries provided by a human operator in the form of either speech commands or text directly. The primary task of this module is to extract a sequence of map search operations and robot actions that need to be performed;
- **Low-Level Planning (LLP)** subsystem — abstracting robot hardware and the specifics of the action implementations from the other modules, this technology block is concerned with actually physically executing the sequence of actions commanded by the HLP;
- **Semantic map** subsystem — an open-set semantic map fusing multiple observations of an environment, such as the workspace of an industrial robot, or the surroundings of a mobile manipulator, into a single, consistent map, where semantic search operations can be performed.

For testing and validation purposes, it is necessary to define the use-case and human operator interaction procedures. The human operator overseeing the robot control system issues commands in voice or text — free form natural language. The robot workstation computer hosts the LLP application which includes the user interface, as the LLP is required to directly integrate with the low-level hardware drivers. Action primitives expose servers implementing the ROS action

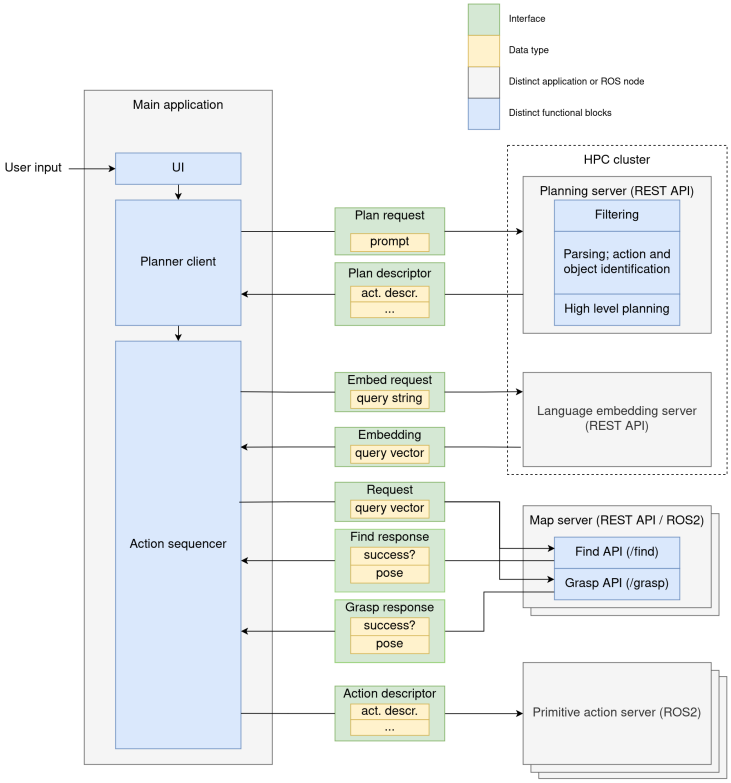

Figure 3.1: Architecture of the industrial robot tabletop manipulation demonstrator, which was built for map-HLP-LLP integration and map validation in the object detection and grasp pose estimation tasks [10, 12].

protocol, performing a preemptable, parameterized motion when invoked. Plans from the HLP and object information from the map are obtained through request-response protocols to servers which may be deployed locally or on physically separate servers. A schematic of the demonstrator architecture as actually built [10, 12] is depicted in Figure 3.1. Real-world data collection and end-to-end execution was performed on EDI's testbed industrial robot workstation. This consists of two robot arms — a *Universal Robots UR5e* and a *UR5*, of which the former is used in this demonstrator; the *UR5e* is equipped with a finger-type *RobotiQ* gripper; above the primary work area, a *Zivid One+* RGB-D camera of the structured light type is mounted statically; a *Intel Realsense D435* is mounted on the robot arm itself. The primary work area is a flat steel surface, with attachment points for boxes that are used in bin-picking and sorting tast demonstrations. As

one of the cameras is static, while the other is mounted on the robot arm — and the robot ROS driver computes tool center point coordinates at high frequency by solving forward kinematics — the usage of SLAM or other separate localization algorithm is not required for building maps in the tabletop manipulation scenario, allowing the semantic map to be tested separately.

The High-Level Planning (HLP) subsystem transforms incoming user requests — voice or text, stated in natural language — into an action plan. To generate the plan, successive calls to an LLM are made, which progressively fill in the details in a set of prompt templates, providing details extracted from the query text itself where required. The HLP uses LLama3 8B Instruct [55] with *Q_4_K_M* weights from [56], implemented in *llama.cpp* [57], with 4-bit GGUF quantization. In addition to the use of quantization to reduce memory requirements, prompt chaining [58] is used to decrease the necessary context length. The HLP subsystem implements three input processing stages: filtration, parsing and planning. The filtration stage classifies incoming requests into one among a set of predefined classes — *"mobile_manipulation"*, which triggers the execution of subsequent planning-related stages in the pipeline, and *"query_answering", "casual_discussion"* or *"any_other_query"*. The next stage in the pipeline, assuming planning has been requested, is instruction parsing. It identifies objects and actions in the user's request, then breaks the request down into simpler subtasks or "steps" for separate planning as in Least-to-Most [59]. The final stage is step-by-step planning. Each step is interpreted in the planning stage as a list of the actions available to the LLP, which are concatenated into the full plan, passed on to the LLP subsystem.

The Low-Level Planning (LLP) subsystem processes plan descriptors — sequences of actions — generated by the HLP. It interfaces with the controller and planner interfaces exposed by robot and other hardware drivers, using *ROS 2* framework for task scheduling and inter-process communication. For motion planning, the ROS2 version *MoveIt* [13] is used, which abstracts the complexity of turning motion goals in Euclidean space to sequences of configuration space targets. A collection of five action primitives has been implemented: **embed**, **find**,**get_grasp**, **activate_gripper** and **move**. All of these are implemented through the ROS 2 service call interface, and treated as composable subroutines when forming the higher level abstraction — **actions**. As of the writing of this document, two actions are implemented in the LLP subsystem — **pick** and **place**.

3.2. Experimental Assessments

To measure and verify HLP subsystem performance, the approach taken in this work is starting with a set of known plans, obfuscating them into natural language commands, then testing whether the HLP, when presented with the obfuscated command, is able recover the structure of the original plan — the sequence of actions and the correct objects associated with them. Given the very

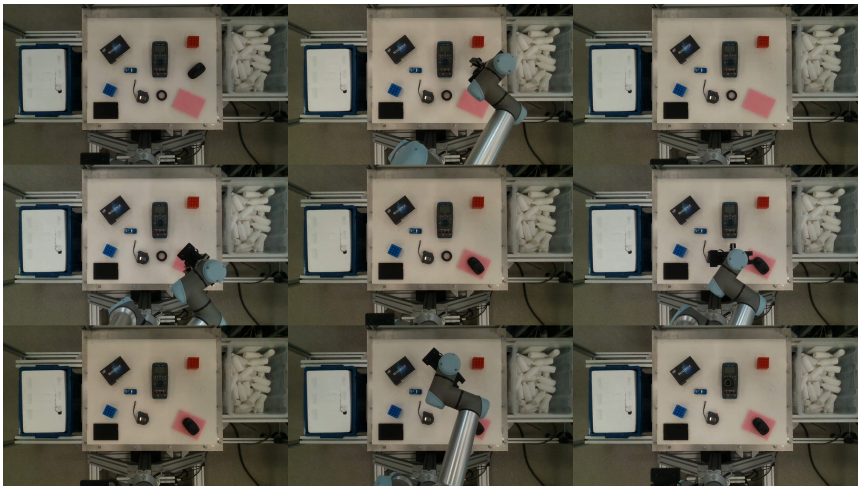

Figure 3.2: Time-lapse of a single command being performed, from the video recording made during one of the end-to-end execution tests. The command given was *Rest the electrical tape on the multimeter, but only after positioning the computer mouse over the pink foam*, which was executed successfully. The robot arm clears the workspace for collecting each new observation, to avoid obstruction.

simple set of actions available to the version of the demonstrator that was tested — the "pick" and "place" pair — three types of command were designed to test for different types of failure that may be encountered during the planning process: "simple"— direct pick-place sequences; "forward"— repeated pick-place commands in order of appearance; "reverse"— two pick-place commands that have to be executed in reverse order. To generate phrasing permutations in two stylistic types, the interactive assistant API of *OpenAI's GPT-4o* [60] was employed. Specific object descriptions can then be inserted into the templates to yield commands.

For end-to-end integration testing and validation a protocol was devised for executing sequences of instructions from a randomized set and scoring the results based on qualitative observations. A time-lapse of one such execution episode is depicted in Figure 3.2. A random permutation of the product of these objects is generated, and a set of instruction templates is sampled. Due to the large number of unknown unknowns involved in a physical integration test, a human observer is present to monitor each episode and note the presence errors in each system stage — HLP, perception, low-level robot control — and assign a score — success, partial success or failure. The total score is given by the minimal score attained in any of the three system stages — a single partial success makes the result a partial success at best, and a single failure means the total is also a failure.

Table 3.1: Robot control system evaluation results.

HLP Verification		
Command type	Count in data set	Correct %
simple	394	99.49
forward	400	92.50
reverse	400	96.40

End-to-end Execution (15 episodes)	
System stage	Success %
HLP	93.33
Perception	86.67
Robot control	73.33
Total	66.67

3.3. Results

The most salient results of the two types of experimental evaluations have been summarized in Table 3.1. The top block in the table records the HLP assessment results. The leftmost column specifies the command structural type (including both stylistic types). The middle column specifies how many times this type of command was represented in the test data set. The discrepancy in types by command count is caused by duplication — the "*simple*" command only has two object slots, meaning some sequences that differ in their last two elements result in the same filled template. The rightmost column lists the percentage of correctly recovered plans for each command type. Upon inspection of the HLP performance evaluation results, it becomes clear that the approach used is quite capable at dealing with the rather limited space of commands represented in the evaluation data, with success rates above 90% for all plan signatures. However, a somewhat counter-intuitive pattern is observed in the fact that reversed plans have lower error rates. Human examination of the specific failed commands reveals that most failures are due to either superfluous actions being hallucinated or object descriptions being duplicated, both of which occur more frequently with the forward ordering. Another prominent issue is the misidentification of polite requests as casual conversation or question answering.

The bottom block in the table summarizes the success rates attained in the latest end-to-end execution test of the entire system. The left column specifies the system stage, and the right column cites the percentage of episodes where this stage succeeded in its task, with the total success rate (when all stages succeeded and the task was accomplished) recorded at the bottom. While it is clear that the system is fundamentally capable of accomplishing simple stacking tasks, future work in improving the performance in each stage is required to make it ready for industry deployment — such as more robust grasp pose estimation.

4. SEMANTIC PERCEPTION SYSTEMS

At the core of the research described in this dissertation was the development, testing and real-world validation of open-set semantic perception systems. All the theses are ultimately concerned with the topic — Thesis 1 concerns experimental methodology used in evaluating such systems, Thesis 2 asserts that an open-vocabulary segmentation model can successfully be used in fine-grained robotic manipulation tasks and Thesis 3 seeks to establish that, through appropriately selected lookup vectors, the vision-language embeddings can be used to distinguish between different terrain types relevant in autonomous robot navigation. Three different types of experimental evaluations were performed, related to each of the theses — the novel reference pose measurement method was used in assessing SLAM system localization performance, the perception systems deployed on the tabletop demonstrator were tested on object recall, while the latest version of the outdoor mapping system — *SLAMVDB* — was evaluted in terrain segmentation accuracy.

4.1. Implementations

As part of an iterative development process, a total of four perception pipelines were developed, in increasing order of complexity, each implementing lessons from the previous steps:

1. The **Depth Map** — a "2.5-dimensional" map, in the form of a single depth image and corresponding vision embedding feature map, used as the initial proof of concept for object detection and grasp pose estimation methods.

2. The **Vector Octree** — a single, fully 3-dimensional octree data structure, holding VLM embedding vectors in its grid cells. Developed primarily to validate methods for integrating multiple semantic observations of the same grid cell, and establishing the transferability of object detection methods used in a single depth map to a fully 3-dimensional map;

3. The **Sequential** Semantic SLAM system — a large scale mapping pipeline exploring ways to adapt the voxel grid map to the challenges faced in outdoor perception — namely, sensor drift accumulation, post-integration merging through loop closures and having to go beyond the volume limits of an integer-indexed voxel cube — through a sequential submap data structure where voxel maps are implemented using sparse matrix indexing;

4. The Time-indexed Semantic SLAM system (**SLAMVDB**) [11] — the final (as of writing this document) version of the perception system, fully decoupling the localization and voxel mapping stages to enable modularity, combining the Vector Octree data structure with the submap approach from

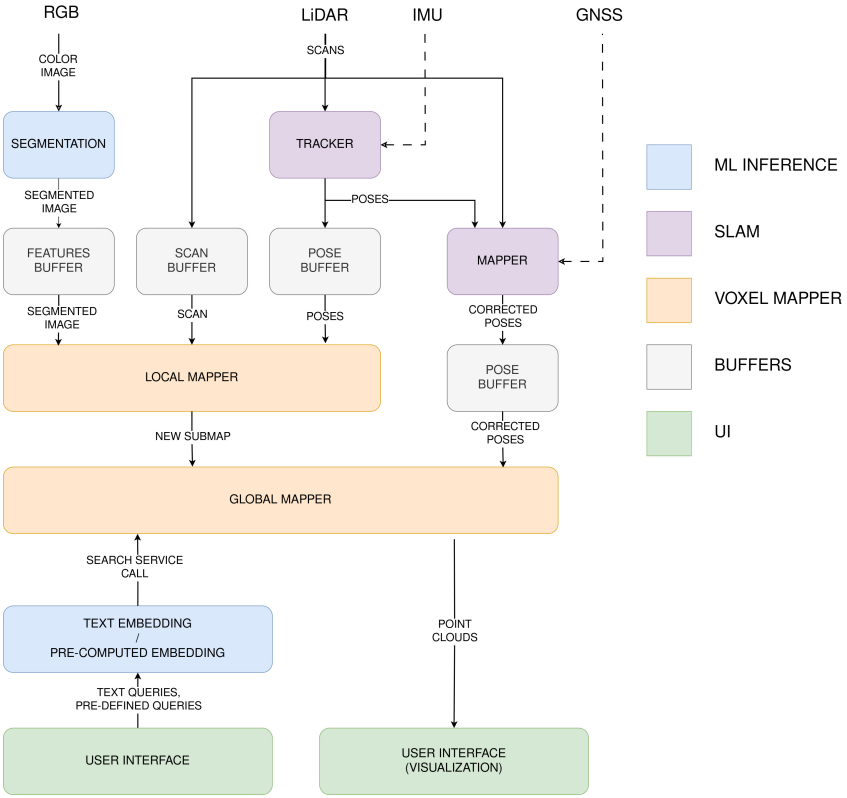

Figure 4.1: Components and data flows in the latest iteration of the open-set semantic mapping system as of the writing of this document [11]. The dashed lines denote optional data flows (IMU for use in external SLAM modules or gravity estimation, GNSS as an optional feature in the mapper).

the Sequential system. A block diagram of this system is depicted in Figure 4.1.

The Depth Map and Vector Octree systems were deployed and tested in the tabletop manipulation scenario [10]. The object detection tests in [10] were performed using these systems, and the algorithms that the Vector Octree was tested with have been directly ported to both subsequent map versions. The Sequential system was used in the initial round of outdoor tests and is featured in the localization accuracy evaluation in [12] — perhaps most importantly, in validating the GNSS-enabled mapping approach. *SLAMVDB* combines lessons learned from the previous iterations and serves as the proof of concept for integrating third-party SLAM systems at the localization stage of the pipeline.

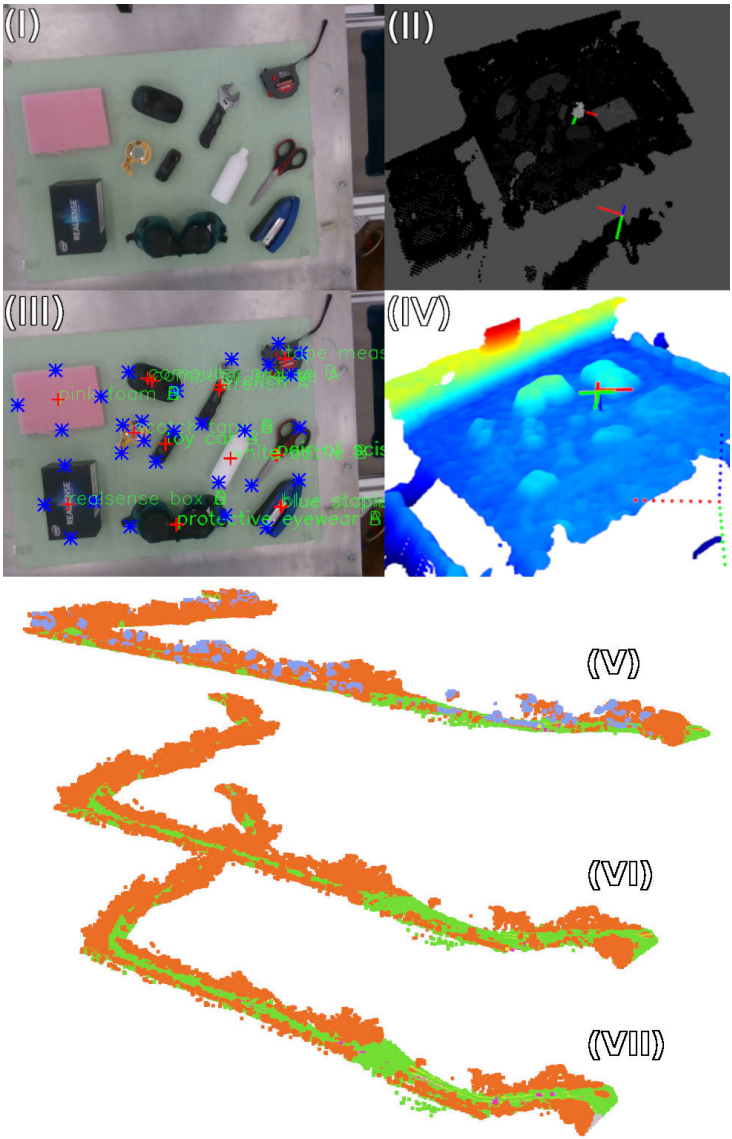

Figure 4.2: (I - IV) A scene in the reference data set. (V-VII) Impact of ignoring "*void*" in the outdoor mapping system.

4.2. Experimental Evaluation

The Depth Map and Vector Octree were evaluated on a reference scene data set as part of the tabletop demonstrator in [10]. To assess the capability of the Depth Map and Vector Octree in finding the desired object given a text description, a reference scene data set was collected, depicted in Figure 4.2 items (I-IV). A scene is arranged (I), and an image of it, loaded into an application specifically developed for the purpose (II), is manually tagged by human annotators for object centroids (red cross markers), grasp directions (blue asterisk markers) and text descriptions (in green). A series of depth images are also collected, which can be reconstructed into a voxel map (III), and object detection can be performed using similarity thresholding (light color in the image indicates similarity to the query "scotch tape"). In evaluation, the map is queried for each object using its text description, and the estimated object centroid or grasp pose is compared to the nearest ground truth pose assigned to the object (IV). To determine whether embedding quantization can be used to reduce memory requirements, tests were conducted at ternary, byte and 32-bit float resolution for the semantic vector scalars.

To verify the quality of the localization estimates used in the outdoor maps, SLAM system accuracy was measured w.r.t. the Apriltag system in the EDI courtyard (a looped trajectory) and an open field (an open-ended trajectory, corrected by GNSS), which was done with the Sequential mapping system. This follows the same general procedure as GNSS-INS pose error estimation detailed in Section 2.2 — though only the ATE [61] metric is considered in the results. The *lio2* [20] tracker supported by *SLAMVDB* was assessed on *RELLIS-3D*.

Finally, *SLAMVDB* was been evaluated in terrain segmentation on *RELLIS-3D*, which provides its ground truth terrain segmentation in two forms — image space masks and segmented LiDAR scans, both of which are sensor-local, and need to be projected into the map reference frame for any comparison. Since even a minor amount of localizer drift rapidly hides any semantic segmentation error, localization accuracy and map construction are treated as separate problems. Thus, for semantic map evaluation the mapping system uses ground truth label poses, and the same trajectory is also used to project the sensor-local ground truth semantic labels into the map frame.

4.3. Results

Table 4.1 collects highlighted results from the three experimental evaluation campaigns. The top row block summarizes the object recall accuracy, which, with the exception of one system configuration, is above 90% for both the map versions tested, and shows no clear impacts from embedding quantization. These results are far from conclusive however, as the reference scene data set is quite small. In several instances, the set of correctly and incorrectly recalled objects is exactly the same — with differences in the final score being due to a handful of

Table 4.1: Results from perception system accuracy assessments.

Object Recall for Grasping		
System	Quantization	Recall, %
Depth Map	ternary	91.04%
	byte	92.29%
	float32	92.29%
Vector Octree	ternary	94.69%
	byte	87.60%
	float32	93.85%

Localization				
System	Data set	Track	l, m	ATE, m
Sequential, loops	*EDI-SLAM*	*courtyard_gt*	696.09	0.057
Sequential, GNSS	*EDI-SLAM*	*saga_gt*	421.86	0.485
SLAMVDB, lio2	*RELLIS-3D*	*00000*	329.76	2.12
SLAMVDB, lio2	*RELLIS-3D*	*00003*	256.58	6.08
SLAMVDB, lio2	*RELLIS-3D*	*00004*	228.41	0.91

Terrain Segmentation Accuracy				
Track	with *void*		without *void*	
	ρ_{hit} %	acc. %	ρ_{hit} %	acc. %
00000	80.28	84.69	81.26	87.35
00003	79.57	79.19	80.39	81.37
00004	73.17	74.22	73.90	76.17

deviations between the various configurations.

The middle row block summarizes results from the localization accuracy tests. On *EDI-SLAM*, the LiDAR-ICP SLAM system clearly demonstrates its ability to counteract drift even along relatively long trajectories — through loop closure in the long, loopy courtyard track, and using GNSS constraints on the open field without loops. On *RELLIS-3D*, even the more complex LiDAR-inertial *lio2* system accumulates a significant amount of drift over shorter tracks, due to the lack of clear tracking features in many segments of the track and rapid rotations of the robot platform — clearly indicating the need to separate localization and semantic accuracy assessments.

The bottom block shows the point-wise classification accuracy ρ_{hit} and voxel-wise classification accuracy attained by the system. Elimination of unobservable "*void*" cells in the map by construction improves accuracy, up to as high as 87.35% on the longest, but least complex *00000* track. This is also shown in Figure 4.2 — partially translucent treetops in (V) get initially mistaken for the sky, but after removing "*void*" from the ontology, their ground truth (VII) "*obstacle*" class is recovered (VI).

SUMMARY AND CONCLUSIONS

In this dissertation, three main theses were proposed and validated — regarding SLAM system localization accuracy evaluation methodology, the use of open-set semantics in fine-grained robotic manipulation in response to natural language commands given by non-specialist human operators, and performing terrain segmentation for UGV navigation by re-using an open-vocabulary image segmentation model through optimizing a set of query vectors. In all three cases, the assertions in the theses have been backed up by experimental evidence. The terrain segmentation method is, to the best of author's knowledge, entirely unique in its approach — having only been proposed as part of the broader open-set perception and human interaction technology developed in [12]. Moreover, the conclusions drawn from the theses have already been re-used in the very same research — the proposed ground truth pose measurement method, the advantages of which over GNSS-INS localization in environments with disrupted satellite signal has been demonstrated, was directly used in experiments SLAM performance verification for the outdoor mapping systems.

Taking into account the results discussed above, new questions and future research directions have become apparent. After discussions with industry representatives, a clear need for more than just single-event driven control for industrial robots has been identified. Rather than the command-action architecture presented in this dissertation, there is a market need for a language-to-program paradigm, where repeatable, adaptable routines are generated from text or voice descriptions. Focusing in on the perception aspects, ways to integrate better-informed grasp and placement pose estimates into the open-set semantic perception system are clearly required, as even when most plans get generated correctly, and the right objects found, the final step of physically manipulating the objects is responsible for the largest fraction of failed execution episodes. In terrain segmentation, novel vision embedding model architectures need to be explored, to overcome the resolution-inference time tradeoff faced in this work. Furthermore, research is required regarding ways to model the presence of mobile objects in the voxel maps — which currently only support holding static objects, with any ephemeral observations being simply cleared. Finally, the need for integrating semantic information into the geometric map construction and localization steps is becoming increasingly pressing — hopefully, reducing the incidence of issues such as *void*-valued occupied voxels, and aiding in SLAM aspects such as loop closure detection.

BIBLIOGRAPHY

1. Forum, W. E. *The Future of Jobs Report 2025* tech. rep. (World Economic Forum, Geneva, 2025). https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf (2025).
2. Ryan, M. *Labour and skills shortages in the agro-food sector* OECD Food, Agriculture and Fisheries Papers 189 (Organisation for Economic Co-operation and Development (OECD), Paris, 2023). https://doi.org/10.1787/ed758aab-en (2025).
3. Arents, J. & Greitans, M. *Smart Industrial Robot Control Trends, Challenges and Opportunities within Manufacturing* in (2022).
4. Badue, C. S. *et al.* Self-Driving Cars: A Survey. *ArXiv* **abs/1901.04407** (2019).
5. Oliveira, L. F., Moreira, A. P. & Silva, M. F. Advances in agriculture robotics: A state-of-the-art review and challenges ahead. *Robotics* **10,** 52 (2021).
6. Oliveira, L. F., Moreira, A. P. & Silva, M. F. Advances in forest robotics: A state-of-the-art survey. *Robotics* **10,** 53 (2021).
7. Racinskis, P., Arents, J. & Greitans, M. Constructing maps for autonomous robotics: An introductory conceptual overview. *Electronics* **12,** 2925 (2023).
8. Wang, J. & Olson, E. *AprilTag 2: Efficient and robust fiducial detection* in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2016).
9. Racinskis, P., Krasnikovs, G., Arents, J. & Greitans, M. The EDI Multi-Modal Simultaneous Localization and Mapping Dataset (EDI-SLAM). *Data (2306-5729)* **10** (2025).
10. Racinskis, P., Vismanis, O., Zinars, T. E., Arents, J. & Greitans, M. Towards Open-Set NLP-Based Multi-Level Planning for Robotic Tasks. *Applied Sciences* **14,** 10717 (2024).
11. Peteris Racinskis. *SLAMVDB - the EDI SLAM Vector Data Base (pre-release version)* https://github.com/edi-administrator/SLAMVDB. Accessed: 2025-05-03.
12. *RoLISe T4.1 Project Report* https://www.edi.lv/RoLISe_T4_1. (Last accessed: 30.04.2025).
13. Open Robotics. *moveit - Package Summary* http://wiki.ros.org/moveit. Accessed: 2025-05-02.

14. Open Robotics. *move_base - Package Summary* http://wiki.ros.org/move_base. Accessed: 2025-05-04.

15. *ROS Wiki: Movebase Global Planner* [Online; accessed on 03-June-23]. http://wiki.ros.org/global_planner.

16. Thrun, S., Burgard, W. & Fox, D. *Probabilistic robotics* ISBN: 9780262201629 (MIT Press, Cambridge, Mass., 2005).

17. Dellaert, F. & Kaess, M. Factor Graphs for Robot Perception. *Foundations and Trends® in Robotics* **6,** 1–139. ISSN: 1935-8253. http://dx.doi.org/10.1561/2300000043 (2017).

18. Sun, K. *et al.* Robust Stereo Visual Inertial Odometry for Fast Autonomous Flight. *IEEE Robotics and Automation Letters* **3,** 965–972 (2018).

19. Bloesch, M., Omari, S., Hutter, M. & Siegwart, R. *Robust visual inertial odometry using a direct EKF-based approach* in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2015), 298–304.

20. Xu, W., Cai, Y., He, D., Lin, J. & Zhang, F. FAST-LIO2: Fast Direct LiDAR-Inertial Odometry. *IEEE Transactions on Robotics* **38,** 2053–2073 (2022).

21. Campos, C., Elvira, R., Rodr'iguez, J. J. G., Montiel, J. M. M. & Tardós, J. D. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE Transactions on Robotics* **37,** 1874–1890 (2020).

22. Qin, T., Li, P. & Shen, S. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Transactions on Robotics* **34,** 1004–1020 (2018).

23. Lu, G., Yang, H., Li, J., Kuang, Z. & Yang, R. A Lightweight Real-Time 3D LiDAR SLAM for Autonomous Vehicles in Large-Scale Urban Environment. *IEEE Access* **11,** 12594–12606 (2023).

24. Dellaert, F. & Contributors. *borglab/gtsam* version 4.2a8. [Online; accessed 01-June-2023]. https://github.com/borglab/gtsam.

25. Hornung, A., Wurm, K. M., Bennewitz, M., Stachniss, C. & Burgard, W. OctoMap: an efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots* **34,** 189–206. https://api.semanticscholar.org/CorpusID:8655888 (2013).

26. Chatila, R. & Laumond, J.-P. Position referencing and consistent world modeling for mobile robots. *Proceedings. 1985 IEEE International Conference on Robotics and Automation* **2,** 138–145 (1985).

27. Kuipers, B. *Modeling Spatial Knowledge* in *International Joint Conference on Artificial Intelligence* (1978).

28. Kuipers, B. The Spatial Semantic Hierarchy. *Artif. Intell.* **119,** 191–233 (2000).

29. Kirillov, A., He, K., Girshick, R. B., Rother, C. & Dollár, P. Panoptic Segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 9396–9405 (2018).

30. Radford, A. *et al. Learning Transferable Visual Models From Natural Language Supervision* in *International Conference on Machine Learning* (2021).

31. Jatavallabhula, K. M. *et al.* ConceptFusion: Open-set Multimodal 3D Mapping. *ArXiv* **abs/2302.07241** (2023).

32. Li, B., Weinberger, K. Q., Belongie, S. J., Koltun, V. & Ranftl, R. Language-driven Semantic Segmentation. *ArXiv* **abs/2201.03546.** https://api.semanticscholar.org/CorpusID:245836975 (2022).

33. Jiang, P., Osteen, P. R., Wigness, M. B. & Saripalli, S. RELLIS-3D Dataset: Data, Benchmarks and Analysis. *2021 IEEE International Conference on Robotics and Automation (ICRA),* 1110–1116 (2020).

34. Wigness, M., Eum, S., Rogers, J. G., Han, D. & Kwon, H. *A RUGD Dataset for Autonomous Navigation and Visual Perception in Unstructured Outdoor Environments* in *International Conference on Intelligent Robots and Systems (IROS)* (2019).

35. Guan, T., Kothandaraman, D., Chandra, R. & Manocha, D. GANav: Group-wise Attention Network for Classifying Navigable Regions in Unstructured Outdoor Environments. *ArXiv* **abs/2103.04233** (2021).

36. Hu, Y. *et al.* Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis. *arXiv preprint: arXiv:2312.08782* (2023).

37. Wang, L. *et al.* A survey on large language model based autonomous agents. *Frontiers of Computer Science* **18,** 186345 (2024).

38. Song, C. H. *et al. LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models* in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2023).

39. Lin, K., Agia, C., Migimatsu, T., Pavone, M. & Bohg, J. Text2Motion: from natural language instructions to feasible plans. *Autonomous Robots.* ISSN: 1573-7527. https://doi.org/10.1007/s10514-023-10131-7 (2023).

40. Ahn, M. *et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances* in *Conference on Robot Learning* (2022).

41. Brohan, A. *et al.* RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *ArXiv* **abs/2307.15818.** https://api.semanticscholar.org/CorpusID:260293142 (2023).

42. Cremona, J., Comelli, R. & Pire, T. Experimental evaluation of Visual□Inertial Odometry systems for arable farming. *Journal of Field Robotics* **39,** 1123–1137 (2022).

43. Rosinol, A., Abate, M., Chang, Y. & Carlone, L. *Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping* in *2020 IEEE International Conference on Robotics and Automation (ICRA)* (2020), 1689–1696.

44. Liu, Y. *et al.* Accurate 3-D Semantic Segmentation of Point Clouds for Intelligent Vehicles Based on Multiview Edge Guidance and Fusion. *IEEE Sensors Journal* **24,** 26853–26865. https://api.semanticscholar.org/CorpusID:270807942 (2024).

45. Schubert, D. *et al.* The TUM VI Benchmark for Evaluating Visual-Inertial Odometry. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* 1680–1687 (2018).

46. Ouster. *OS1 Hardware User Manual* https://data.ouster.io/downloads/hardware-user-manual/hardware-user-manual-revd-os1.pdf. Accessed: 2025-05-08.

47. Basler. *daA1920-160uc* https://docs.baslerweb.com/daa1920-160uc. Accessed: 2025-05-08.

48. Movella. *MTi-680G* https://www.xsens.com/hubfs/Downloads/Leaflets/MTi-680G.pdf. Accessed: 2025-05-03.

49. Furgale, P. T., Rehder, J. & Siegwart, R. Y. Unified temporal and spatial calibration for multi-sensor systems. *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems,* 1280–1286. https://api.semanticscholar.org/CorpusID:15778738 (2013).

50. Burri, M. *et al.* The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research.* eprint: http://ijr.sagepub.com/content/early/2016/01/21/0278364915620033.full.pdf+html. http://ijr.sagepub.com/content/early/2016/01/21/0278364915620033.abstract (2016).

51. Racinskis, P., Arents, J. & Greitans, M. *Annotating SLAM data sets with Apriltag markers* in *2024 10th International Conference on Automation, Robotics and Applications (ICARA)* (2024), 438–442.

52. Terzakis, G. & Lourakis, M. I. A. *A Consistently Fast and Globally Optimal Solution to the Perspective-n-Point Problem* in *European Conference on Computer Vision* (2020). https://api.semanticscholar.org/CorpusID:226239551.

53. Creative Commons. *Attribution-NonCommercial-ShareAlike 4.0 International* https://creativecommons.org/licenses/by-nc-sa/4.0/. Accessed: 2024-11-01.

54. EDI. *EDI-SLAM data* http://edi.lv/EDI-SLAM_dataset. 2024.

55. Dubey, A. *et al. The Llama 3 Herd of Models* 2024. arXiv: 2407.21783 [cs.AI]. https://arxiv.org/abs/2407.21783.

56. Bartowski. *Llamacpp imatrix Quantizations of Meta-Llama-3-8B-Instruct* https://huggingface.co/bartowski/Meta-Llama-3-8B-Instruct-GGUF. Accessed: 2024-10-22.

57. Gerganov, G. *llama.cpp* https://github.com/ggerganov/llama.cpp. Accessed: 2024-07-14.

58. Wu, T., Terry, M. & Cai, C. J. *AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts* 2022. arXiv: 2110.01691 [cs.HC]. https://arxiv.org/abs/2110.01691.

59. Zhou, D. *et al. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models* 2023. arXiv: 2205.10625 [cs.AI]. https://arxiv.org/abs/2205.10625.

60. OpenAI. *GPT-4o System Card* https://cdn.openai.com/gpt-4o-system-card.pdf. Accessed: 2024-11-14.

61. Sturm, J., Engelhard, N., Endres, F., Burgard, W. & Cremers, D. A benchmark for the evaluation of RGB-D SLAM systems. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems,* 573–580 (2012).