**Robotics: Science and Systems 2024**
**Delft, Netherlands, July 15-July 19, 2024**

# Demonstrating Event-Triggered Investigation and Sample Collection for Human Scientists using Field Robots and Large Foundation Models

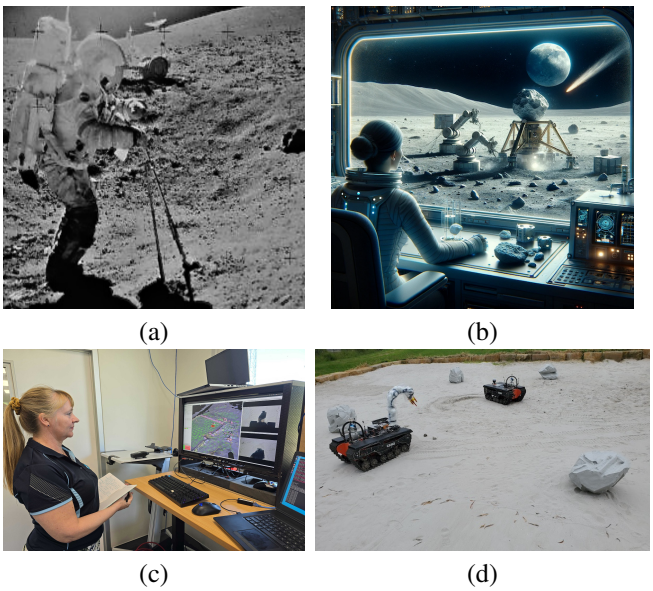
Tirthankar Bandyopadhyay*, Fletcher Talbot*, Callum Bennie*, Hashini Senaratne*, Xun Li*, Brendan Tidd*, Mingze Xi*, Jan Stiefel*, Volkan Dedeoglu*, Rod Taylor*, Tea Molnar*, Ziwei Wang*, Josh Pinskier*, Feng Xu*, Lois Liow*, Ben Burgess-Limerick*‡, Jesse Haviland*‡, Pavan Sikka*, Simon Murrell*, Jane Hodgkinson†, Jiajun Liu*, Fred Pauling* and Stanislav Funiak*
*CSIRO Robotics, Data61, †CSIRO Mineral Resources, ‡QUT Center of Robotics
Brisbane, Qld 4069 Australia
Corresponding author: tirtha.bandy@csiro.au

***Abstract*—In this paper, we introduce a pioneering end-to-end system demonstrated on a team of robots and sensors, designed to augment scientific exploration and discovery for human scientists in remote or inaccessible environments. We demonstrate and analyse our system's capability in a mock-up test-bed scenario. In this futuristic hypothetical scenario human scientists located in a controlled lunar habitat, are assisted by a team of robots in investigating an unknown seismic phenomena like moon-quakes or meteor impact detected by a sensor network deployed on the lunar surface. They do so by autonomously collecting data, providing contextual semantic information and collecting scientific sample for future analysis upon the direction of humans. This work is among the earliest to present a feasible way to integrate large foundational models (LFMs) into field robotic deployment, enabling easy semantic and contextual understanding of the objects in the environment and natural language-based interactions with the robot for the scientist. In addition we bring together state-of-the-art techniques in mapping, object detection, navigation, mobile manipulation, soft grippers, event detection and present details of the integration, insights and lessons learnt from the deployment. While demonstrated in a limited setting in a mock-up environment with ground robots, the system architecture and approach presented in this paper is easily generalised with domain-specific customised components and robots for a variety of event-driven scientific discoveries e.g., geological survey, biodiversity study or underwater environmental sampling.**

***Index Terms*—Field Robotics, Event-detection Sensor Networks, Scene understanding, Mobile manipulation, Human robot interfaces, Foundation models**

## I. Introduction

Scientific data collection from the field is one of the most critical first steps that enables scientists to study and solve complex problems and reach groundbreaking discoveries. Emerging AI-integrated robotic technologies can play a crucial role in providing scientists with important assistance in scientific field data collection. In this paper, we present a demonstration of a field robotic sample collection system that takes human input and performs an autonomous human-in-the-loop investigation and sample collection task in the space exploration domain.



(a) (b) (c) (d)

Fig. 1: Our motivation for this demonstration paper is to present results on our analysis of moving away from human heavy sample collection for scientific discovery to robot-aided scientific sampling. (a) Shows Apollo 16 astronaut John W. Young collecting lunar samples with a scoop (credit: NASA). The vision of a human scientist focusing on higher level scientific analysis while directing a fleet of robots in the field to perform sample collection or exploration is captured in an artistic image in (b) (credit: OpenAI). Our work mocks up such a vision in a concrete demonstration where human scientist in (c) interacts with a fleet of remote robots from a base station giving it natural language commands. The two robots used in the demonstration, Explorer and Collector, are shown working together in the bounded sandpit amidst mock-up boulders in (d).

### A. Demonstration Scenario

For this demonstration, we created a futuristic hypothetical scenario of a fleet of robots operating on the lunar surface, with the human scientist performing experiments in a lunar habitat. Sensors akin to NASA's Apollo-era lunar seismic detectors, placed to study the sub-terranean structure via measurement of moon-quakes and meteor strikes are expected to be deployed [23]. Our futuristic scenario goes further in imagining a situation where the scientist (either on Earth or on a lunar base) is able to investigate a seismic phenomenon due to the availability of a fleet of heterogeneous robots on the lunar surface. Upon being informed of an unknown seismic event, the scientist without the need for doning a spacesuit herself to investigate and collect samples (like prior Apollo missions Fig.1(a), interacts with an AI driven autonomous system to send specific robots for closer inspection and sample collection. An artist's imagined scenario is captured in Fig.1(b). Our demonstration mocks up this scenario using existing and novel system components like the base-station Fig.1(c) and a fleet of field deployable robots and sensor networks Fig.1(d).

The scenario of lunar seismic event based investigation was selected for its encapsulation of the most formidable challenges inherent in both robotic deployment and scientific exploration. While on one hand it addresses the open scientific questions surrounding extraterrestrial and lunar geology, it also exposes the inherent nature of these operations demanding significant reliance on autonomous robotics due to the logistical and environmental constraints of extraterrestrial settings. Furthermore, this demonstration tackles the critical issue of optimizing valuable human crew efforts, which are often expended on non-scientific tasks, thereby enhancing the efficiency and scientific yield of these missions [34].

### B. Impact of this demonstration study

When space scientists collaborate with carefully designed, appropriate AI-enabled robotic technologies, they can benefit from delegating tedious physical processes to agile and robust robots without being exposed to harsh conditions and also collect samples from hostile environments that they cannot reach by themselves [35]. Involving robots on-board decision-making abilities in sample collections also increases the precision of sample collection (e.g., in tracking the location of collected samples) [15] and speeds up the process of documenting samples with contextual information (e.g., using automatically captured real-time video/audio) [15]. Furthermore, AI-based naturalistic interactions and other AI-based features incorporated into those robotics systems can enable scientists to arrive at well-informed decisions efficiently by optimising their time and attention on their capacity to analyse robot-communicated complex information and make decisions based on their domain expertise and guide robots in real-time for collecting productive and appropriate samples [35]. Using robots to navigate risky environments and physically collect samples, while receiving supervisory support from scientists to choose appropriate samples, reduces the physical strain and cognitive load on scientists. This is in comparison to to trials that engage robots by overloading the human with teleoperation control of the robot [26, 24] or other low-level technical duties, resulting in situational awareness latencies on other important tasks [36]. Therefore, this approach allows scientists to best use their time on higher-priority tasks like performing scientific analysis or experimentation that require their domain expertise. Further, since this approach optimises the usage of complementary skills of scientists and robots, extending the capabilities of both parties and achieving sample collection missions that they cannot complete alone [11]. In turn, it expedites scientific research and exploration.

While we present a mock up scenario of an event aligned with a lunar rover, the systems approach and the underlying components are widely applicable to a variety of other domains, including marine sampling of sea grass [3], biomaterial sampling for bio-diversity studies or rare/novel species detection using DNA analysis, event based detection of the spread of pests or diseases in an agricultural field among many others.

Our system expands prior work in presenting an end-to-end pipeline that specifically focuses on human scientific exploration. This is achieved through the active involvement of our scientists in interacting with the robots, allowing them to gather semantic contextual information from the area of interest and leverage state-of-the-art foundation models. Additionally, we are showcasing this demonstration using robust field-deployed robots. As a result, our framework allows us to seamlessly integrate a combination of higher-level reasoning and low-level perception and control.

This demonstration paper presents the following main contributions:

- Presents an end-to-end autonomous system for event-driven human scientific exploration, investigation and discovery, enabling human scientists to best use their domain expertise skills.
- Demonstrates a feasible way of integration of large foundational models (LFMs) in the loop of field robotic deployment, enabling easy semantic understanding of the objects in the environment and natural language-based interactions with the robot for a remote scientist.
- Presents learning on integration challenges and insights into systems development, event detection, semantic visual mapping, field mobile manipulation, navigation and human-robot interaction during the deployment that would be extremely valuable to the scientific community to design better systems in the future.

## II. System Description

### A. Robotic and Sensing hardware

The autonomous system used in this deployment consists of seven vibration sensors, two robot agents and a base-station computer used for operations. The two robot agents are both customised BIA5 OzBot All Terrain Robots (ATR) [17] designated, in this paper, as Explorer and Collector which additionally has a robotic manipulator arm. Both Explorer and Collector made use of CSIRO CatPack perception payloads [13],

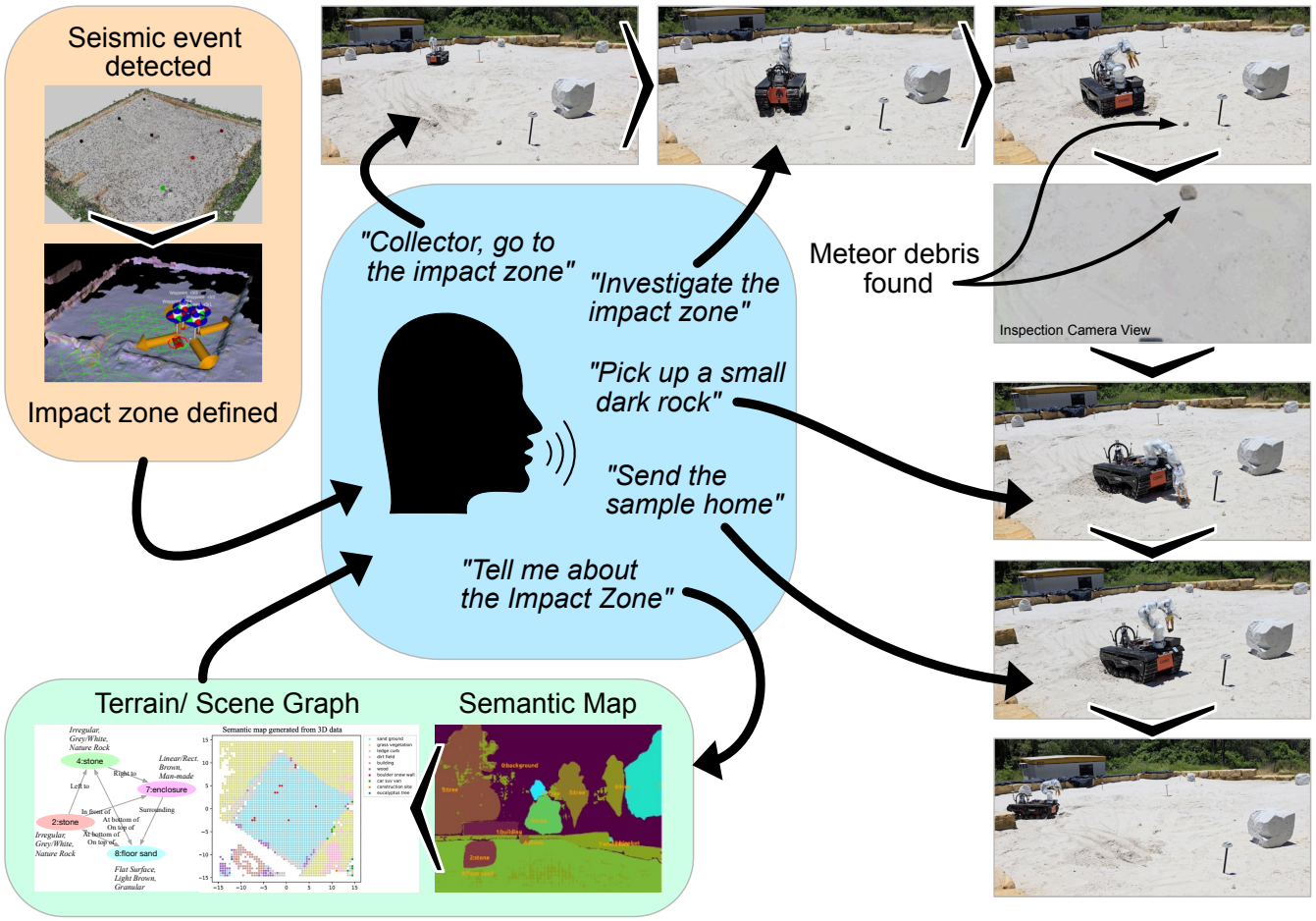



Fig. 2: This figure describes the overall pipeline of the demonstration performed. In the offline phase the robots explore the area of interest and impact sensors are placed. Upon detection of the impact event, the system interacts with the human using natural language to decide on the tasks to be performed. The robots then perform specific tasks autonomously at the scientist's request. A video of this demonstration is included in the supplementary material.

which run Wildcat SLAM software, and separate autonomy compute systems running CSIRO NavStack software, a multi-agent autonomy and navigation solution. Collector also was fitted with a Franka Emika Panda 7 axis robot arm, which was used for mobile manipulation and sample collection tasks.

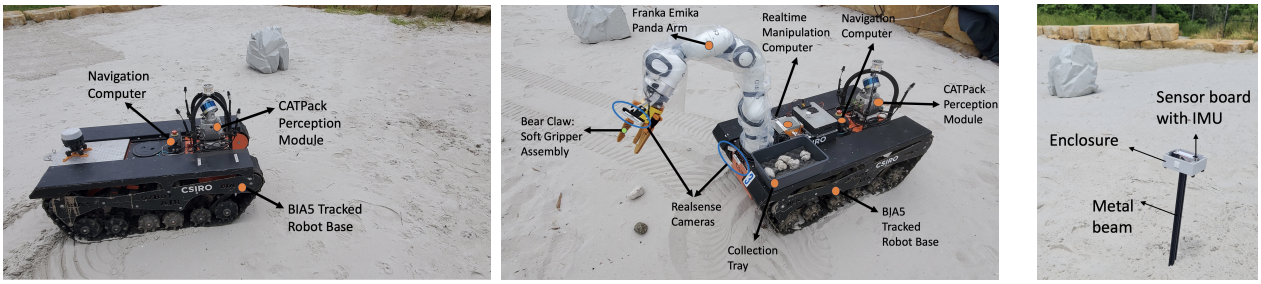
The user base-station was networked to both robot agents using a communications mast installed with a Rajant Breadcrumb Peregrine node and a Rajant Breadcrumb ES1 node installed on each agent. The Rajants provided a layer 2 mesh network using wireless (2.4 GHz and 5.8 GHz) 802.11 radio links [17].

Sensor networks can facilitate event-driven sample collection missions by continuously monitoring the mission area for detecting events of interest. Once an event is detected by the sensor network, robots can be guided to optimal sampling locations using the estimated location of the event.

To identify optimal sampling locations triggered by an impact event, a sensor network consisting of 7 sensor nodes designed to sense vibration was deployed on the sandpit as illustrated by Figure 3(c). Each sensor node has an XIAO Sense board with the Inertial Measurement Unit (IMU) LSM6DS3TR-C and is powered by a battery. To better capture the vibration signal in the sand, the sensor node hardware amplifies the signal using mechanical amplifiers. The base of the sensor is a wooden plate measuring 45cm x 30cm buried 20cm under the sand. The wooden plate is intended to capture as much of the vibration signal as possible. A metal beam, 50cm in length, is mounted orthogonally on the base and protrudes from the sand. The base transfers the vibration onto the beam, which works as a lever amplifier. At the top of the beam, a watertight enclosure houses the electronics. The sensor is mounted on a lever amplifier in the enclosure to further amplify the analog signal.

### *B. Multi-Robot Autonomous Navigation and Exploration*

The mutli-robot navigation system fielded in this demonstration, called NavStack, is built upon capabilities developed during the DARPA SubT challenge where fleets of robots




(a) Explorer robot configuration (b) Collector robot configuration (c) Vibration sensor node

Fig. 3: Physical robot and sensing hardware components. More lower level component details can be found in [13]. Note that while the robots had GPS module, it was not used for localization and navigation.

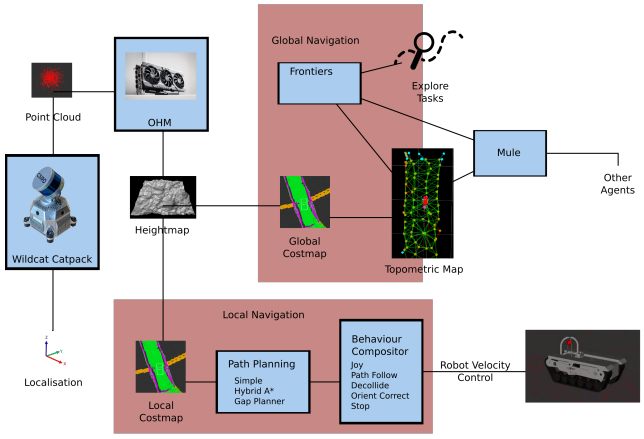



Fig. 4: Data flow pipeline used in local navigation module for CSIRO NavStack

explored and mapped subterranean environments [17]. Key capabilities of the DARPA SubT system reused were local navigation, autonomous exploration, topological mapping, global navigation, executive control and monitoring. An improved path planning algorithm was added to the system in order meet constraints set by the lunar exploration scenario. The local navigation functions eases human workload by automating obstacle detection and avoidance. Autonomous exploration reduces cognitive workload by facilitating time efficient exploration coverage of the region of interest. Topological mapping and global navigation enabled an efficient return to the region of interest by the Collector robot and traversal between the points of interest. Reducing cognitive load on robot operators allows them to manage more tasks therefore reducing the number of humans required. In a lunar exploration setting, requiring fewer people reduces the cost to conduct experiments. Efficient path planning is also important due to the limited time and energy budget robots have on the moon.

Central to NavStack is the CSIRO WildCat Simultaneous Localisation and Mapping (SLAM) software. WildCat software executes directly on the CSIRO CatPack perception hardware producing local odometry, multi-agent global localisation solutions and processed lidar sensor readings localised into the odomotery solution's frame of reference. In this application, the CSIRO NavStack software uses WildCat exclusively as inputs to its navigation and autonomy calculations. The Graphics Processing Unit (GPU) based Occupancy Homogenous Map (OHM) software tool is employed by NavStack to process lidar and localisation data creating an occupancy map and heightmaps. Special to the height maps provided by OHM are heightmap layers which include Autonomous Ground Vehicle (AGV) specific three dimensional mapping data, allowing AGVs to plan paths over overpasses and under underpasses. NavStack processes heightmaps into cost maps which are used by a local navigation software to move robots to goals set by the executive software. Local navigation consists of a hierarchy of path planners that perform navigation and obstacle avoidance within the immediate area of the robot and a selection system to choose the most suitable planner for the navigation goals presented. NavStack can receive multiple control inputs, including from the path planners, and uses a prioritized behaviour compositor to determine which input has priority to actuate the robot. The global navigation solution is a graph based representation of simplified traversability information that represents a topometric map (Topomap or Topo). Topomap is built on the mutlit-agent global localisation solution provided by WildCat and Topomap data can be shared amongst the robots on the system either by saving for reuse later or real-time sharing using the communication layer protocol developed for the DARPA SubT challenge (codname "Mule"). The global navigation solution also detects areas yet to be visited (frontiers) converting them to exploration tasks so that the robot can return to unexplored areas. Finally, at the highest level of the system, the executive software receives navigation way points or high level tasks through a human machine interface or programatically from other systems. Figure 4 illustrates the organisation and flow of data through the subsystems of NavStack.

Due to the limited scale and scope of the scenraio presented, only exploration using a single AGV has been demonstrated. Based on CSIRO's prior work in DARPA SubT, more Explorers can be added including multi-modal AGVs (wheeled, tracked, legged) as well as aerial platforms. The task bidding system used in the DARPA SubT challenge [25]

will be employed to efficiently distribute exploration tasks between agents For subsequent Collector operations a new task can be introduced and again the task bidding system used to efficiently distribute the tasks amongst agents. As an extension to the task-bidding system, multi-point route planning software can be created to optimise the tasks won by individual agents. A knapsack style minimum cost, maximum value algorithm can be implemented for task prioritisation or potentially a Travelling Salesman Problem heuristic solver, such as Christofedes algorithm [6] can be used for scenarios where locations of all detections must be visited.

Navigating on granular materials is challenging both on the traversability challenges as well as local controllability issues. Our tracked vehicles are able to surmount the granular slippage due to their higher area of contact and did not get stuck in the terrain. However the goal tolerance and body posing was often not sufficient to perform a pick. This requires the development of a local search planner to bridge the gap between the coarse navigation goals with the fine body positioning required for sample collection.

### C. Semantic Understanding Pipeline

Semantic mapping is a process of building a map that contains the semantic meaning of the objects in the environment. The constructed semantic map helps the robots understand the environment better, in addition to the colour and geometry captured from the sensors. The map can benefit robotics navigation and route planning by providing cues of the objects in the scene. The semantics of objects can also provide prior knowledge for the manipulator to choose the optimised approach to interact with and grasp the target objects. Finally, the semantic embedding space empowered the search and reasoning for complex queries. As illustrated in the Figure 5, the semantic label generation process utilises pre-trained Large Foundation models (LFM), which significantly reduces the human effort to annotate a large amount of data and are able to be generalised to various use cases. During the semantic labelling process, the raw images captured from the robot camera are first processed with off-the-shelf open-set auto-tagging model, Recognize Anything Model (RAM [42]), to provide abundant textual annotations for every image. Next, the grounding model GroundingDINO [21] is adopted to locate the objects that appeared in the images. The grounded object bounding boxes are then forwarded to a zero-shot segmentation model Segment Anything in High Quality (SAM-HQ [16]) to obtain fine-grain segmentation masks. The semantic patches of the given image are then encoded by a cross-modal encoder CLIP [29] into embeddings, hence the vision and language embeddings are in the same feature space for cross-modal search. The semantic masks from 2D images are projected into the 3D space using lidar measurements and an estimate of the robot pose. Finally, the projected embeddings associated with the semantic masks in a 2D voxel grid are summed and renormalised; this procedure amounts to estimating the mean direction of a von Mises-Fisher distribution generating the samples in each 2D voxel.

The semantic voxel map stores detailed information about objects and their surroundings. For illustration, an example semantic map is shown in Figure 6. In each grid of this voxel map, an accumulated 768-dimensional CLIP embedding is stored to capture the visual features of that location. During the process of locating specific objects at inference time, e.g., "big white rocks", the module extracts the CLIP embedding of the given query and compares it with the stored embeddings in the semantic map. The Figure 6(c) demonstrates that the areas matching the query are confidently highlighted in yellow, while the surrounding sand surfaces are coloured with dark shade. In addition to testing semantic mapping on the lunar testbed ($20\times20m$), large-scale mapping was also conducted on CSIRO's QCAT site, which spans approximately $500 \times 600m$, as shown in Figure 7. The mapping results provide a comprehensive overview of the semantics of the entire site, and clearly demonstrate the scalability of the proposed semantic understanding pipeline.

In addition, drawing inspiration from vision-language grounding [22] [18] [4], we further infuse a richer layer of semantic information into our mapping process, which will form a scene graph. This approach diverges from conventional object detection and segmentation by placing a stronger focus on the attributes of objects and their spatial interrelations, articulated through natural language. More specifically, by employing pre-trained LLMs (with vision capability) and prompt engineering techniques, we create associations for each object in a 2D scene with its distinct attributes, including shape, color, size, and material. In these 2D environments, we also extract and analyze the spatial relationships between adjacent objects. This rich semantic layer enhances the context-awareness of natural language interactions, facilitating precise entity localization and object manipulation. The goal is to enable our system to not only recognize objects but also comprehend their broader contextual significance, leading to more intuitive and effective human-robot collaborations.

The current deployment for LFMs and semantic mapping pipelines is running on a desktop machine equipped with a Nvidia GeForce RTX 3090 GPU ( 36 TFLOPs, 24GB vRAM). As for perception pipeline, a Jetson Xavier AGX was used to process the point cloud data. Despite the current deployment being on a desktop machine, it is feasible to adapt the main component for more efficient deployment. There are efficient implementations for the components used in the semantic mapping. For example, in the segmentation pipeline, the powerful segmentation module SAM has a few lightweight counterparts such as MobileSAM [41]. The smaller MobileSAM effectively replaces the original ViT-H encoder (632M parameters) with a significantly smaller Tiny-ViT (5M parameters) that is capable of operating on edge AI devices.

### D. ImpactSense: Event Detection

Distributed sensors provide an additional layer of information as part of the holistic representation of the physical world to facilitate the robot's interaction with it. In this demo, Inertial Measurement Units (IMU) are used to detect impact event and

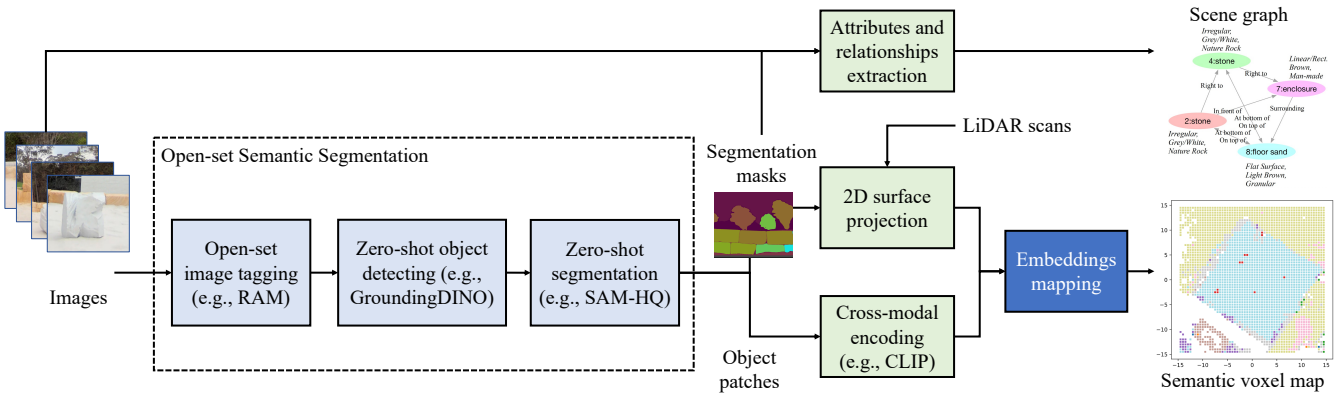



Fig. 5: Workflow of offline semantic map generation. In this pipeline, the 2D semantic masks are firstly extracted from the images. The 2D segmentation masks are then projected into 3D using LiDAR scans. Given the 2D-3D alignment, the semantic embeddings are mapped into the corresponding voxel grids. A scene graph is further extracted to capture the attributes and relationships of detected objects.

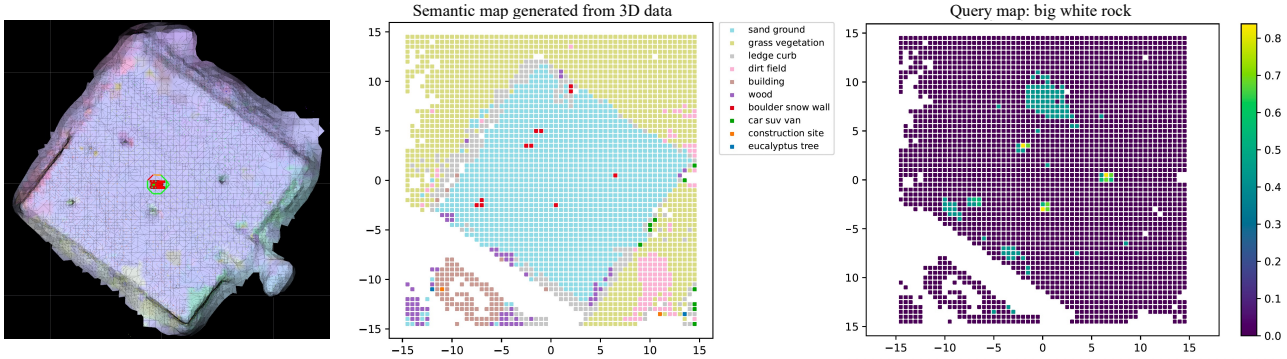



Fig. 6: Visualisation of semantic mapping. (a) 3D mapping of the sandpit from the top-view angle. (b) A visualisation of the voxel semantic map, coloured by the most-related label from the text corpus. Each grid is encoded with a 768-dim CLIP embeddings. (c) A result query map given the query "big white rock" (highlighted in yellow)

subsequent robot presence. In real-life, this contributes to the tracking of the robot's trajectory and behaviour for validation and safety purposes, and to the optimisation of task planning for the robots.

To detect impact events of interest, a threshold-based time-domain vibration analysis method was used. The threshold was calibrated to be sufficiently high to discern a simulated impact event while ensuring that routine activities (such as the movement of rovers and robots) would not be mistakenly identified as impact events. The locations of the sensor nodes detecting an event are shared and an *impact-zone* is identified and sent to the autonomous robotic navigation system for local region inspection and sample retrieval. For seismic event detection based on the time-domain analysis of the vibration signals, we use a detection threshold of 0.2g. Based on the application requirements, constraints, and sensing capabilities, different event-detection methods can be utilised (see [8] for a review of event detection in sensor networks.

It is worth noting that the same principle can be applied to other types of sensors as well as sensor-based events, such as sudden temperature/humidity change from temperature/humidity sensors, or presence detection through infrared/radar-based presence sensors. As a generic process, sensor data streams and events are then fed into the LLM/planner to optimise the robot's task planning specific to the use case.

### *E. Natural language interface with Robots*

An interface based on natural language enables users to interact with complex systems with minimal prior knowledge or specific training using the platform. Large language models in the interaction pipeline also unlock a level of complex reasoning and semantic understanding that is otherwise very challenging to encode. This section describes the natural language pipeline used to interface with the various subsystems in this deployment, facilitating a seamless interaction with a domain expert unfamiliar with the system interface.

The natural language pipeline in use was initially designed for search and rescue operations within the SubT challenge and then repurposed for the presented "Seismic event" mock-up

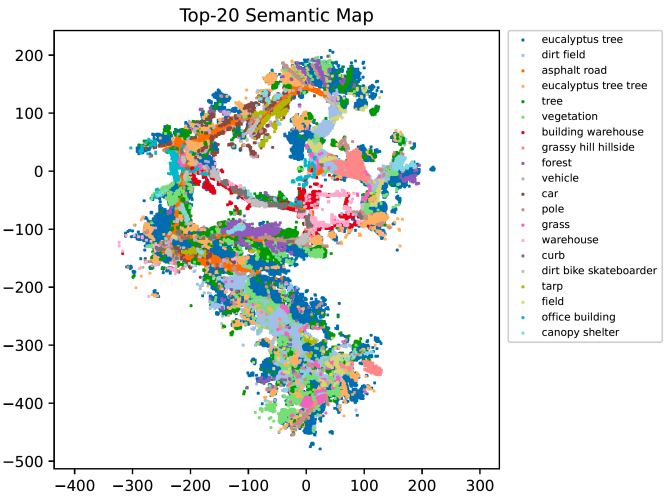



Fig. 7: Visualisation of a large-scale semantic map of the CSIRO's QCAT site, coloured by the most-related label from the text corpus. Each point is encoded with a 768-dim CLIP embeddings.

demonstration with further improvements. In search and rescue scenarios, operators are trained extensively and require a level of trust in the autonomous operations of the system [32]. Through the SubT challenge [17], a graphical interface was developed for a single expert human supervisor to oversee a fleet of robots in subterranean environments. The interface offers several levels of autonomy from teleoperation to full out-of-communication autonomy [5]. However, language has emerged as a preferred modality for interacting with robots, particularly for users unfamiliar with the interface, providing system status with plain language queries and enabling advanced reasoning afforded by large language models [30, 37, 19]. For our SubT fleet we developed a natural language interface, named "Squawk" [1], consisting of a speech-to-text interface, GPT module with a prompt to interpret user speech input and output formatted code that can be executed by robot.

The speech-to-text interface uses the `large-v3` model from the Whisper API[1]. Written in C++, the speech interface runs comfortably on a mid-range laptop. The wake word "Hey Squawk" initiates dialogue with the system using a model provided by Picovoice[2]. Alternatively, the user can initiate an interaction from a keyboard press. Once a query has been parsed, the response from the GPT query is displayed on the screen and converted to audio using the Piper text-to-speech model[3].

The Squawk interface builds on work using LLMs that take user queries and outputs formatted code that can be executed for robot control [37]. We use GPT-4[4] with a crafted prompt for in-context learning that contains sections including an assistant role, available functions, 3D scene graph and example interactions (Fig. 9). The assistant role is informative of the persona to be adopted and conditions GPT on the expected performance for the task. The functions GPT can call are described in typical Python syntax and are commented accordingly. The scene graph and available agents are inserted into the prompt at the time of inference and describe which robots are active as well as semantic labels and their associated poses. Figure 8 shows the key user requests for the demo outlined in this paper and the responses generated by GPT. The generated functions are parsed within a ROS1 Python system and the associated actions are executed on the robot fleet. More details of Squawk can be found in Bennie et al [1].

The interface provides full access to the multi-agent navigation system, including initiating full-autonomy exploration operations and navigation to, and storing semantic landmarks. The user also has full access to the system information of all deployed robots and can perform intuitive GPT queries from the same interface. Whilst the available functions provided to GPT are simple, they can be utilised by GPT in interesting ways. For example, simply commanding "Gather all robots at Explorer's location" can send an arbitrary number of robots to Explorers position. Commands such as "Have Collector patrol in a 5x5 square" will chain together commands to complete this geometric based path.

For the demonstration presented in this article, the robots support an astrogeologist in the collection of rock samples (Fig. 14(c)). Through Squawk, the domain scientist can specify commands to the robot team in a flexible and intuitive way, without experience and with minimal guidance. First, the Explorer robot is instructed to go to the operational region and provide a coverage map of the area. With the shared map, the Collector robot is instructed to go to impact zone, as identified by the ImpactSense component, to investigate and collect a specific sample type (e.g. a small dark rock). The scientist can at any time query the status of the mission or discuss science outcomes with a GPT instance.

### *F. Local Region Investigation*

Upon receiving a command "Go to Impact Zone" from the scientist, the collector autonomously proceeds to the impact region. Upon reaching the region the robot is able pose its wrist camera towards the region of interest to investigate and provide closeup images to the scientist. This view coupled with the semantic information of the region available to the scientist allows her to decide on whether to collect a sample of interest.

Upon recieving a request for sample collection, locating the sample for collection occurs in two stages. Once the robot has navigated to a pose near where a sample is expected to be found, an initial search is performed. The manipulator targets the wrist-mounted camera at the point of interest, and the object detector is queried. When a target sample has been identified, fine localisation is initiated. The robot uses the holistic control method described in Section II-G to move the camera to a pose $0.6\,\mathrm{m}$ directly above the identified

[1]https://platform.openai.com/docs/models/whisper
[2]https://github.com/Picovoice/porcupine
[3]https://github.com/rhasspy/piper
[4]https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

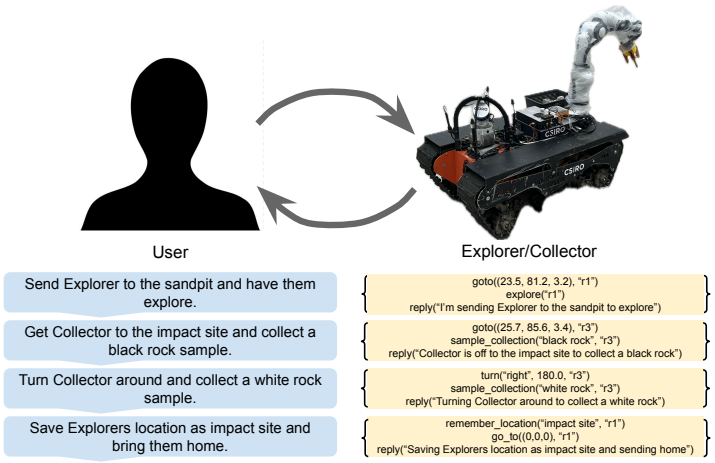



Fig. 8: System overview with sample user requests and GPT responses.

**Assistant Role:**
You are an intelligent and friendly virtual assistant.
You are helping a user control a fleet of field robots.
…

**Active Robots:**
{current_agents}

**Available Functions:**
# Use Text to Speech to reply to the user with informative and useful content
reply(reply: str)
# Send a robot to a position (x, y, z) tuple.
This should be inferred from the User request and the 3D Scene Graph
go_to(position: tuple, RID: str)
…

**3D Scene Graph:**
{scene_graph}

**Object Class Count:**
{observed_object_count}

**Examples:**
…

You are an assistant that uses Python code to perform actions and respond. Use the list of available functions to make function calls to best serve the users request. You must only respond with a sequence of function calls.

**User:**
{user_input}

Fig. 9: Summary of the GPT prompt

sample. The object detector is queried again which confirms the identification and refines the position estimate.

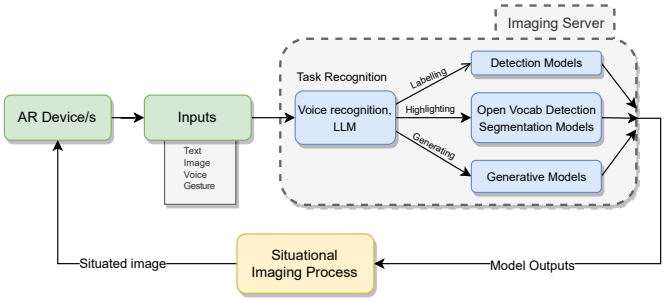



Fig. 10: An illustration of the original Situated Imaging pipeline [40]. The SI pipeline is capable of processing multi-modal inputs and perform multiple downstream computer vision tasks simultaneously.

Upon receiving the scientist's instruction, such as "pick up the small dark rock", the robot performs the rock detection utilising an imaging pipeline derived from the Situated Imaging (SI) pipeline [40], which is an extensible array of techniques for in-situ interactive visual computing. An illustration of the original SI pipeline is shown in Figure 10. In this demonstration, instead of augmented reality (AR) devices, the "Input" came from both a human user (voice) and robot (frames). Squawk acted as the "LLM/Task Recognition" component that extracts the entities from the scientist's instruction (i.e., small dark rock in this case). GroundingDINO [21] model, an open-set detector, was used in this particular demonstration to detect rocks of different properties, such as shape and colour. Instead of going through the "Situated Imaging process", which is designed to overlay model outputs onto physical object via AR headsets, SI returns the model outputs to the robot as a formatted JSON with bounding boxes and normalised centre points of the detected rock instances. These are then transformed to an estimated 3D pose of the target object in the local scene.

One of the main advantages of SI pipeline is the use of a server-client design that loads and warms up selected chains of models. By keeping the warmed up model in the VRAM (GPU memory), the imaging pipeline responds to the client (e.g., a robot or a human) requests immediately, leading to greatly reduced latency critical for autonomous real-time behaviours. While we ran the object detection over the network, the detection module could be run on a GPU enabled compute unit onboard the robot.

### *G. Whole-Body Reactive Manipulation Control*

Most mobile manipulation systems perform navigation and manipulation tasks separately – the base remains stationary while the manipulation reaches to grasp an object. As shown in Figure 11, the kinematics of the manipulator on our Collector robot results in a small reachable workspace at ground level. Consequently, it is likely that repositioning of the base will be required to reach an identified object, incurring additional time costs. Significant speed improvements are enabled by controlling both the base and the arm with a single, holistic controller [12].

We adapt the controller presented in [12] which converts desired Cartesian end-effector velocities to coordinated base and joint velocities, while using the redundant degrees of freedom in the system to avoid actuator position and velocity limits and maximise manipulability. This is coupled with a simple position-based servoing loop to perform reaching tasks. The resulting controller is reactive and compensates for disturbances throughout the motion. Reactive capability is particularly important for a tracked vehicle operating on

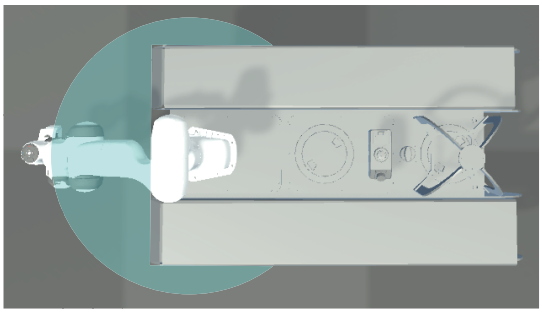


Fig. 11: Top-down view of the Collector robot. The blue shaded area displays the area in which the manipulator can reach and grasp objects off the ground.

unstable and uneven terrain, such as our sandpit, because motion of the vehicle can result in unpredictable sliding and deformation of the terrain. Various disturbances that apply to the Collector robot while manipulating are visualised in Figure 12. The reactive control architecture effectively compensates for these disturbances enabling robust grasping.

### *H. Gripper Design and Grasping Experiments*

Collecting unknown scientific samples in the field vastly increases the complexity of reliable and stable grasp execution compared to conventional pick-and-place. Quality sensing, perception and interactions are all impeded by unknown, dynamic, and time-varying conditions. Visual perception, for both object identification and ground estimation, is rendered unreliable by the variable lighting conditions created both by natural changes in solar luminance and light scattering off the uneven lunar surface. As a result, grippers can easily dig into the loose, sandy lunar terrain and erroneously record a successful grasp when using impedance or closure as a proxy.

Soft grippers overcomes these limitations by shifting control from external digital sensors and processors to their in-built mechanical design and embodied intelligence. Soft gripper are made from flexible materials, which are able to conform

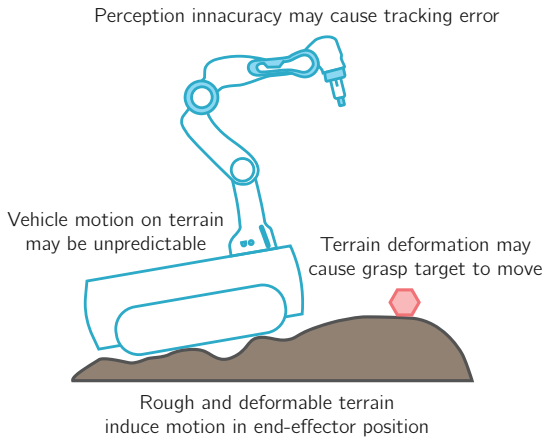



Fig. 12: An overview of the various errors and disturbances that impact the robot's motion as it attempts a grasp.

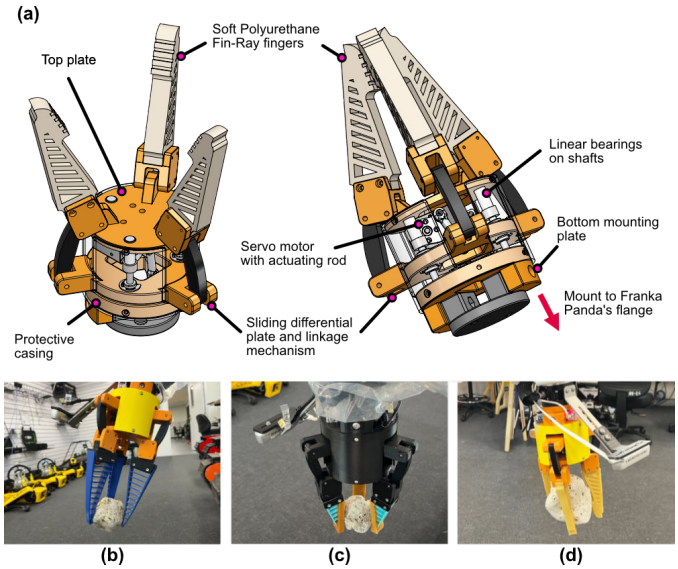



Fig. 13: (a) Mechanical design and assembly of the Bear Claw. (b) First gripper iteration using off the shelf fin-ray fingers from Festo. (c) Second gripper iteration with shorter off the shelf fin-ray fingers with soft fingertips. (d) Third gripper iteration with entirely soft polyurethane fin-ray fingers.

to deformable objects, as well as object with unknown or irregular geometries [14], especially for our rock sample. They can also reconfigure to adapt to their environment [27, 10], a critical capability in unstructured environments [28, 20] and where sensitive physical interactions are required [39], such as ground interaction tasks. The inherent compliance of soft grippers absorbs energy during collisions [33], reducing the need for high frequency feedback and control.

A series of both rigid and soft grippers were experimentally evaluated in both laboratory and field settings. A rigid parallel jaw gripper (the Franka Hand by Franka Robotics) provided a baseline against which to compare our bespoke grippers.

The Franka Hand only consists of two rigid fingers, which lead to two major issues: 1) Precise knowledge of the ground level was required, as the rigid fingers frequently became bogged in soft terrain, and could suffer damage if they collide with hard terrain. 2) Even where the object and ground were accurately localised, a stable grasp was difficult to achieve, as objects tended to eject out of the finger plane during jaw closure. At best, sample rocks were grasped in a small contact patch, resulting in a unstable grasp with poor disturbance rejection.

To address these issues, the 'Bear Claw', a custom gripper shown in Fig. 13 (a), was designed and mounted to the flange of the Franka Emika Panda arm. The gripper utilised a single Dynamixel servo motor (model: XM540-W270-R) with a differential plate to actuate all three fingers. The fingers of the gripper can be interchanged as needed to accommodate different grasping tasks or requirements, as shown in Fig. 13 (b)(c)(d). The servo has an inbuilt PID control and position feedback to enable accurate close-loop control. An Intel Re-

alsense depth camera (D415) mounted to the end of the Panda arm was used to detect and measure objects of interest.

## III. Field Demonstration and Results

(a) (b) (c)

Fig. 14: Demonstration Setup

*Overview of the demonstration:* The demonstration was performed in a Sandpit created at CSIRO's QCAT site, as a representative of the granular lunar terrains as seen in Figure 14(a). The red markers in Figure 14(a) shows the impact location where a heavy rock was dropped manually to simulate meteor-impact. The boulders induce mobility restriction and the vibration sensors are placed to detect the robot's motion and impact detection. Figure 14(b) shows the location of the actual operator base station to the field setup. The operators and scientists in the room had no visual sight to the sandpit and all visualisation feedback and robot commands were happening via a base station over a network. Figure 14(c) shows a professional astrogeologist unfamiliar with the robot system interacting using natural language. The screen show the real-time view of the data stream of the robot cameras. While our communication is robust to failures and bandwidth limitations, we do not yet consider the earth-to-moon latency in the visualisation which might require a different visualisation mode.

Our sandpit is a representation of the environment with granular materials with particle sizes close to lunar regolith and as such captures the essential properties of surface traction, compaction and bearing strength for robot traversability encountered by lunar rovers. Our samples consisted of natural scoria rocks, a kind of volcanic rock, and pumice stones ensuring different physical and visual properties from the granular sand. The boulders were made of gray Styrofoam pieces to allow easy rearrangement for testing obstacle avoidance and visual inspection components.

*Premapping and exploration:* Prior to the impact event, the robot Explorer, autonomously mapped the Sandpit region to create a traversability map and collect pointclouds and image sequences to process the semantic map of the region. 15 shows the pointcloud of the experimental region with the robot exploration trajectory overlaid.

*Enriching semantic map with attributes and relationships using LLM:* To test the capability of attribute and relationship extraction, we provide "GPT-4 with Vision" (gpt4-vision-preview) with both original images and their corresponding 2D segmentation masks from our semantic map generation pipeline as shown in Figure 16. To use prompt engineering to facilitate the output, we follow the best practices [7] to

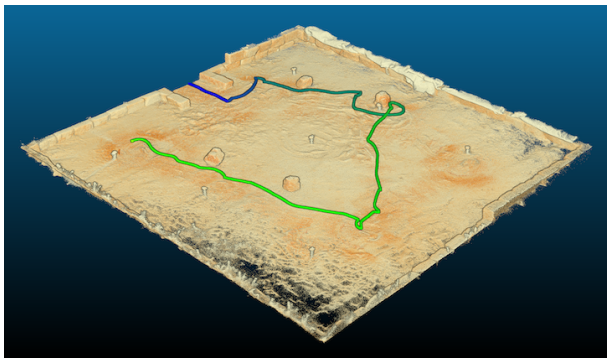


Fig. 15: Autonomous exploration of the operational region by Explorer. The trajectory is overlayed on the terrain map generated from the WildCat SLAM pipeline

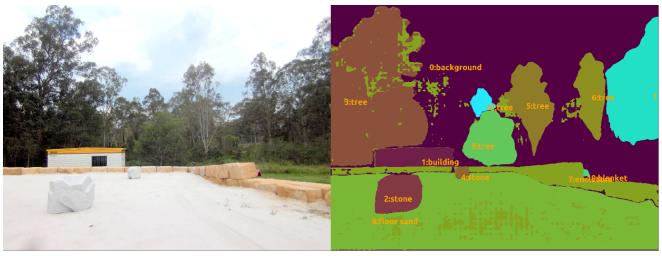



Fig. 16: Example image (left) and its 2D mask for semantic as visual input (right)

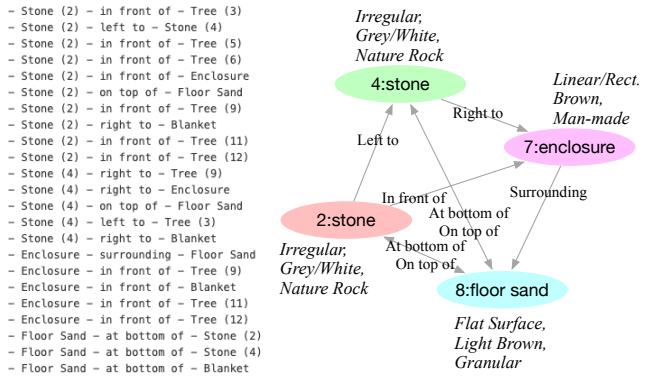
provide clear and specific instructions, incorporate context, explicit constraints, and interactive conversations to guide LLM to generate more accurate and relevant responses, such as background context as "you are planning route and object manipulation for scientific sample collection by the robotic mobile platform. Try to analyse the objects in the given image as part of the mapping process", as well as mandated the 4 types of attributes to be extracted for all the objects. The attributes' output is given in Figure 17and relationships in Figure 18(a). Figure 18(b) illustrates the partially built scene graph of objects within the sample image captured area of the sandpit, highlighting their relationships and attributes.

The intended usage of the scene graph is to allow field robot to furnish field robots with a comprehensive semantic understanding of their environment, enabling context-aware interactions. This scene graph implementation is a work in progress, possessing its own set of limitations that will be discussed in the next section.

**Object Features:**

| Object | Shape | Color | Material | Others |
|---|---|---|---|---|
| Background | N/A | Various (nature) | N/A | Encompasses all other objects |
| Building | Rectangular | Yellow/Brown | Man-made | Single-story structure |
| Stone (2) | Irregular | Grey/White | Natural Rock | Positioned on sand |
| Stone (4) | Irregular | Grey/White | Natural Rock | Positioned on sand |
| Tree (Multiple) | Organic | Green/Brown | Vegetation | Trees surrounding the area |
| Enclosure | Linear/Rect. | Brown | Man-made | Surrounding part of the area |
| Floor Sand | Flat Surface | Light Brown | Granular | Underneath the stones and open areas |
| Blanket | Rectangular | Blue | Fabric | Lying on the sand left to Stone (4) |

Fig. 17: Object attributes extracted from gpt4-vision-preview




(a) (b)

Fig. 18: (a) Object attributes extracted from gpt4-vision-preview, (b) Object spatial relationships extracted from gpt4-vision-preview


Fig. 19: The sequence of images (a-d) capture the robot position and an impact event during the 21-second mock-up demonstration: (a) robot passes by sensor 3 at t=7s, (b) robot passes by sensor 2 at t=12s, (c) a large rock is thrown in to the sandpit to demonstrate a seismic event at t=15.8s, and (d) robot passes by sensor 1 at t= 20s.

*Event detection:* To demonstrate the detection of unknown seismic events, we deploy a wireless sensor network in the sandpit as shown in Figure 19. The network consists of 7 vibration sensors sending vibration data to a BLEAcon node using the Bluetooth Low Energy (BLE) beacon protocol based on a star network topology. The sampling rate for the sensors is set to 104Hz. We use BLE advertising with an advertising interval of 10s for time synchronisation. Note that, for larger sensor network deployments, more energy efficient communication (e.g., multi-hopping [9], where other nodes are used as relays to transmit data) and computing (e.g., in-situ processing [38], where raw data is processed locally at the sensor node) mechanisms can be utilised for improving energy efficiency and network lifetime).

Our event detection demonstration scenario involves a robot moving in the sandpit for 21 seconds while a seismic event happens at t=15.8s triggered by an impact caused by a stone thrown into the sandpit. Figure 19 shows the images captured for the robot position and the impact event during the demonstration. As shown in Figure 20, the vibration

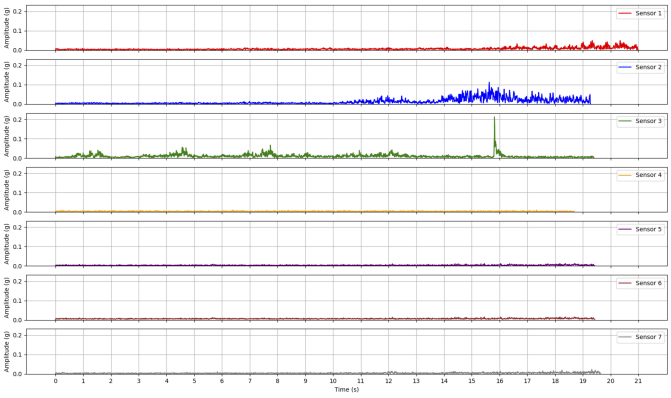



Fig. 20: Vibration signals detected by the sensor nodes during the 21-second mock-up seismic event demonstration.

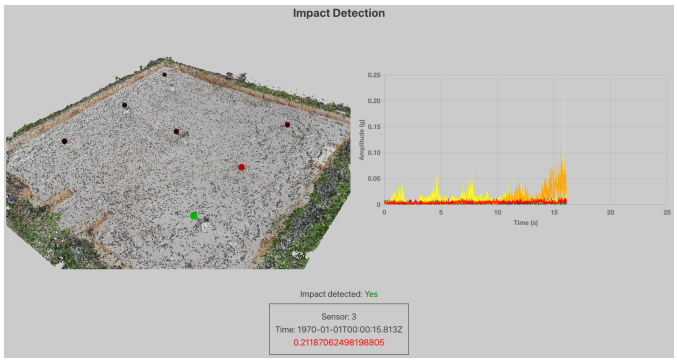



Fig. 21: Real-time impact monitoring application. On the left-hand side, the 3D map of the sandpit with the pre-deployed vibration sensor network is displayed. The colours of the nodes represent the received vibration signal amplitudes. The real-time vibration sensor data is plotted on the right-hand side. The application uses a threshold-based method to detect impacts. When an impact is detected, the application alerts the human scientist displaying the details of the detection.

signatures of the robot movement and the impact event are captured by the sensor network. The vibration sensor data is fed into a real-time impact monitoring application, which detects the impact based on the pre-determined threshold of 0.2g as shown in Figure 21. The application alerts the human scientist by displaying the sensors detecting the event with the measured signal amplitudes and the timestamps. The sensor locations detecting the event are shared with the autonomous robotic navigation system for impact location investigation and sample collection. Note that, accurate localisation of the event zone may require a priori knowledge of signal propagation characteristics and dense sensor deployment.

*Impact location investigation and sampling:* The sequence of images from Figure 22(a-j) shows the whole progression of the Collector robot searching for a ”white rock” and collecting it in its collection box. In Figure 22(b) the robot uses its gripper mounted camera to look at a predetermined search pose to detect the rock in its view. From Figure (c-e)

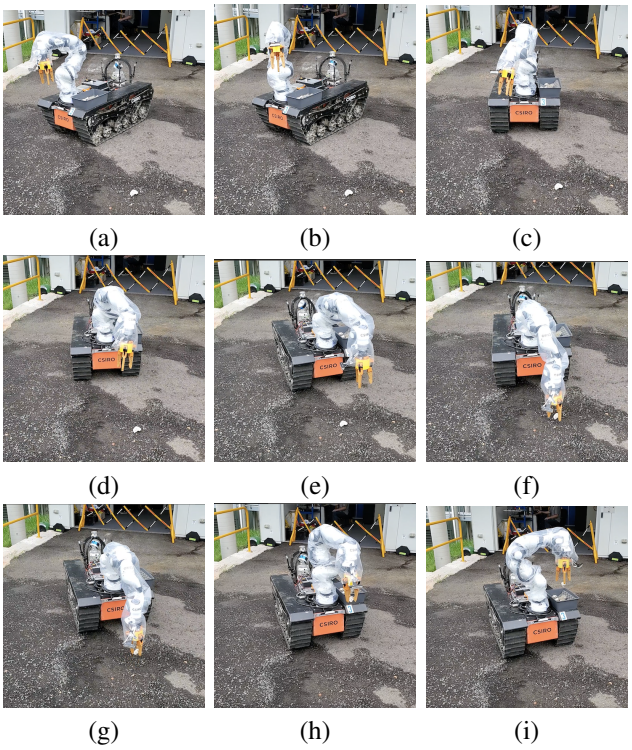


Fig. 22: Snapshots of the Collector going to the impact zone, using its wrist camera to investigate the regions and search for the “small dark rock” as requested via Squawk interface. Upon receiving a positive detection, a the sample is successfully collected.

the robot uses its whole body reactive motion to position itself to approach the rock for picking. At Figure 22(e) the robot poses its gripper camera for a closer look and validate that the sample matches the description given by the scientist. At this stage the human operator could abort the pick and proceed with another search position if the sample is not of value or if additional information is required. Upon confirmation of the sample validity, the robot executes an open-loop pick action Figure 22(f-i) and drops the sample in the receptacle.

**Quantitative Results** To evaluate the sample collection pipeline quantitatively, we ran the pipeline 40 times: 5 times per prompt in two different lighting settings with 4 different object types. The 4 prompts were : pick a ”small dark rock”, ”small rock”, ”red apple” and ”green apple”.

Fig.23 shows the performance results of the experimental run in sample picking. Due to the open set query for sample selection and collection, we were seemlessly able to incorporate objects like ”Red-Apple, Green-Apple” without any modification or tuning of the system. The camera calibration error resulted in wrong positioning of the target pick location leading to 7.5% failure cases. Additionally due to erroneous depth perception from the camera in hand, the arm had overshoot posing resulting in overtorquing the arm. This also includes cases where the gripper made contact with the ground before the sample was caged. The grasping failures came

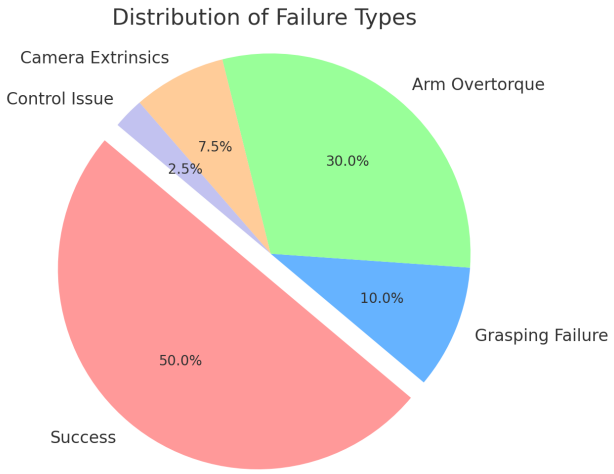



Fig. 23: The pie chart shows the performance results and failure types during the indoor outdoor picks on 4 classes of objects

from the softness of the gripper not able to extract enough gripping force for an sample. The control issues in positioning the whole body and arm were quite minimal. These however did not capture any arm-collision cases as we assumed that the robot would have sufficient space of operations. This assumption is invalidated in cluttered environments and better controller is required.

## IV. Observations, Failure Cases and Lessons Learnt

We present some of our observations on the successes and failures of deployment of the system components during our field tests below.

### *Integration challenges and observations*

One of the biggest lessons learnt from this demonstration was the effort needed to ensure that each individual components work seamlessly with each other towards a successful sample collection. We present below some of the steps we took to ensure reliable system integration and development.

1) *Standardization across platforms with higher level abstraction.* As we demonstrated in our previous work [17], enabling suitable abstraction of the modules decoupled from the lower level system dynamics is key for fleet deployment. This requires standardization of behaviour stack and response rates of various heterogeneous platforms.
2) *Dedicated effort in systems integration testing* A key lesson from software development methodologies is the importance of regular integration testing for all components, whether for experimental behaviors or reliable enhancements. Considerable effort in frequent field deployments ensures subcomponents meet specifications

and exposes feature mismatches and edge cases often missed during the design phase of system architectures.

3) *Field Mobile manipulation as an integration challenge* We encountered a convergence of issues at the systems level—sensing, processing, and representation—that were interconnected in the ostensibly simple task of selecting a conspicuous rock in the sand. Failures in any component, such as the gripper, arm positioning, camera calibration, body positioning estimate, or terrain deformability, hinder accurate sample collection. The results of the sampling attempt are displayed in **??**. This experience underscores the significant challenges involved in outdoor environments, highlighting the need for further research and development to ensure reliability.
4) *Natural language integration* The system parameters are not as flexible as natural language, necessitating either prompt tuning or providing operators with specific keywords to include in their requests. Although we did not encounter significant difficulties with this, overlooking this critical step frequently results in the system rejecting the request. Care has to be taken to ensure that the terminology of the domain keywords of the scientist are contextualized for the robot operation. This is a work in progress for our system.

### *Adaptive or Compliant soft grippers improve the pick rate in field environments*

From the grasping experiments, we found that adaptive grippers were useful for mobile manipulation tasks in uncertain terrains, as they enabled the grasping of objects without the need for high frequency force feedback or accurate positional control. The inherent compliance of the material allowed the gripper to conform to the shape and features of the sampled rock, achieving stable and robust grasps regardless of the disturbances that occurred during the manipulation task. The soft grippers could compensate for positioning errors of the robot arm, whereby the grasped object was off-centred from the gripper and the ground level was higher than estimated, allowing the grippers to dig into the ground without resulting in damage to the manipulator arm. Overall, future research directions should aim to address the challenges in autonomous grasping and manipulation in unstructured environments, where there are large uncertainties due to poor sensing and limited bandwidth. By effectively addressing these challenges, robotic manipulation systems can be utilised with increased reliability to autonomously perform dangerous and remote tasks, and thereby minimising the need for human intervention.

It is important to note that scientific sample collection is often more than just picking and dropping. Proper stowage is a critical component of the collection process. To address this, we intend to work on designing better grippers and receptacles that are robust and efficient in field conditions.

### *Reactive whole body manipulation was critical in handling kinematic limitations and position uncertainty in the field*

In the field as well as in our controlled experimental testing off-field, we found that enabling the whole body integrated motion towards picking samples greatly increased the success rate. Due to the positioning and the kinematics of the arm used, the region of interaction on the ground plane with the arm was quite limited. The base repositioning during the arm ground approach for the pick was critical in avoiding the joint limits and maintaining a high manipulability score. This allowed us to do finer pose corrections than would have been possible with fixed base arm pick motion. The final stage of grasping was an open loop due to camera view constraints. However, a continuous visual servo till the pick could be achieved with an in-hand camera rather than wrist camera as demonstrated in [2].

One key limitation with this control scheme currently is that it does not consider any terrain or obstacle data when executing motion of the base and manipulator. This presents the limitation that the current grasping task can only be executed under the assumption that the environment is approximately planar and unobstructed. Future work to incorporate obstacle avoidance for both the base and the arm would allow such a control scheme to operate in much more complex environments, further broadening the range of potential applications.

### *Semantic mapping provides invaluable contextual information for deciding the parameters of the inspection or sample collection task*

From the semantic mapping experiments, we found that relying solely on a 2D image-based semantic segmentation model does not consistently provide robust results, due to the complexity of the environmental elements. However, by aggregating the 2D observations into a global map, we can estimate the feature embedding at each location more reliably. The proposed approach uses a sequence of pre-trained Large Foundation Models to build the semantic map, which establishes a robust base for inspection or sample collection tasks, while avoiding the resource-demanding local training.

We further observed that the vision-language cross-modal embedding encoder is effective in extracting feature representations of objects and the surrounding environment. The extracted cross-modal embeddings can be searched using both language and image queries, enabling enhanced interaction through Squawk interface.

The promising results from semantic segmentation support the approach of exploiting a comprehensive world representation enriched with detailed information about the objects and their environment. The scene graph illustrates how the attributes and relationships of objects serve as abstracted information for comprehensive semantic understanding. The current limitation mainly comes from the incapability to accommodate viewpoint changes, which is inherent in 2D scene graphs. It struggles to fully capture the depth and spatial relationships necessary for accurate object interaction in a three-dimensional space. To overcome this, we plan to introduce a 3D scene

graph in our forthcoming research. We argue that a hybrid 3D world representation, which integrates abundant object attributes and relationships, is necessary to facilitate more precise and flexible queries for localising and manipulating the objects of interest in a large scene. Furthermore, the generation process of the scene graph will incorporate cross-modal encoding, ensuring a unified representation across both voxel map and scene graph.

***For long-term or repeated deployment of robots in granular surfaces, low impact path planners are highly desired***

Fig. 24: Long-term robot motion causing degradation of the terrain and worsening traversability conditions at the same time making sample collection difficult in the future.

Navigating through granular materials such as sand generally did not present substantial difficulties in linear traversal. However, notable positioning errors occurred during the robot's on-spot rotation, leading to some scene-detection failures. These errors were particularly evident when the sand was compact, a condition frequently observed in post-rain scenarios where layers of wet sand were present just beneath the dry surface. Continuous navigation and repositioning of the base significantly altered the surface contours, in some cases damaging the investigation site before the scientific sample was collected as shown in Figure 24. It became apparent that this type of movement could detrimentally affect not only the navigation surface but also progressively impact the area's navigability.

This navigation challenge provides a unique opportunity to develop innovative local path planning and base positioning algorithms accounting for the interaction with granular materials and aiming to control the dynamic engagement of the robot's base with the surface. The goal would be to approach the target area while minimising terrain disruption.

***Contextual information is required for improved object detection***

We ran a variety of detectors, Yolo8 [31], GroundingDINO [21] and human in the loop-detection where the operator could click on the object to pick in the operator GUI. As expected the operator enabled detection was vastly superior to many of the automated detection but suffered significantly on the pick latency. The current limitations for our system lied in characterising the object for segmentation especially in partial occlusion and similarity to background substrate. While this is expected for zero-shot scenarios (as opposed to in-domain scenarios), several methods can be used to further enhance the zero-shot performance. The most important one is choosing an accurate prompt (i.e., description of the object of interest). Figure 25 shows the significant improvements brought by a good prompt. When using a generic description like "rock", the model picks up all rocks that appear in the frame, including a grey boulder. By adding additional description, such as "small", we narrowed down the predictions to grabbale rocks. Additionally, with the addition of colour description "dark", the model successfully predicted the rock of interest. Additional post-processing, such as analysing colour intensity of detection object of interests, could effectively improve the results. Adding semantic contextual information would vastly improve the detection and local scene analysis for picking.

(a) "Rock" (b) "Small rock" (c) "Small dark rock"

Fig. 25: Detection results using three related but distinct prompts.

***Natural language interface vastly improved the accessibility of robot operation for domain scientists untrained in field robotic operations***

Being able to interact with the robot using flexible natural language by issuing commands and queries made the scientist work with the robots on the sample collection mission without having prior knowledge or training on the conventional visual interface or memorising any robot-specific jargon and constrained language commands. Further, our integration of LLMs and GPT models with web access provided untapped opportunities to extract object relevant information to aid in providing contextual information of the event to better understand the science behind it. While these are not novel ideas, these observations were validated in our experiments. This section outlines some of the challenges and future investigation as language becomes an integral component of our system.

Seamless speech-to-text interaction is difficult to achieve. While there have been rapid advances in the pipeline of speech-to-text in the capability of newer models (including accuracy and reduced latency), there is a degree to which a new user needs to become familiar with the dynamics of the interface. For example, the latency of speech detection and the response from the system initially causes a back-and-forth of partial dialogue as the user is unsure if a command has gone through. With assistants that are now present on most

smartphones, users have an expectation of how a speech-to-text interaction should be conducted, usually with the presence of a visual and auditory cue. For our system, we print direct feedback to a terminal, but implementing *ready-to-listen* and *thinking* sensory cues would improve usability for new users.

Language has undoubtedly improved accessibility to our system for a non-expert user, however, our multi-robot solution is a complex system with many sub-components with various capability. A new user needs to be provided with a detailed description of each component and a list of example commands. The natural language interface requires the user to understand what capabilities are available. Feedback from the language system could improve this interaction by providing the user with a suitable list of commands from the current state. Additionally, mission state and system status information can be provided in an intuitive way without an initial overloading of information. These features are currently being explored.

While we scratched the surface of what is possible with an advanced reasoning agent in the loop with our system, there are many avenues for exploration and expansion. For example, semantic reasoning is one of the capabilities made available by interfacing with a language model, however there are still points of feedback required by the user at each intersection of technology. Work towards an undisrupted system experience is underway. Furthermore, while our user can query large models with the presented setup for general queries (e.g. ask chatGPT a question), we expect end-users would be experts in their own domains. This pushes the question of having separate models for dedicated purposes. The most prominent of these being an assistant targeted at the scientific method, one that is versed in subject matter, can suggest experiments, as well as analyse and visualise data. This exciting concept will be explored in future work.

## V. Conclusion

In summary, this paper has successfully demonstrated a large-scale robot deployment designed to assist human scientists in the field. We detailed an integrated system encompassing the detection, examination, and scientific sampling of unknown events, all within a framework that includes a human scientist in the operational loop.

We have showcased the application of large language models (LLMs) in enabling natural language interactions with the robotic deployment, enhancing the efficiency and effectiveness of these explorations. Additionally, we have provided comprehensive results and discussions, highlighting the potential and current limitations of our approach. The paper has also presented open challenges and identified key areas for development in realising a fully functional, realistic system. We believe that this is just one of the first steps towards realising the vision of rapid scientific discoveries done by team of human and robots.

## Acknowledgments

This paper presents work that was partially funded by CSIRO Robotics strategic funds and partially by research projects namely *Robot teams for off-world exploration* (CSIRO Space and Astronomy Business Unit, CSIRO Data61 Business Unit), *Science Digital* (CSIRO Data61 Business Unit) and *CSIRO Collaborative Intelligence Future Science Platform.*

## References

[1] Callum Bennie, Bridget Casey, Cecile Paris, Dana Kulic, Brendan Tidd, Nicholas Lawrance, Alex Pitt, Fletcher Talbot, Jason Williams, David Howard, Pavan Sikka, and Hashini Senaratne. Alternative interfaces for human-initiated natural language communication and robot-initiated haptic feedback: Towards better situational awareness in human-robot collaboration. In *workshop at 35th Australian Conference on Human-Computer Interaction (OzCHI '23)*, 2023.

[2] Ben Burgess-Limerick, Chris Lehnert, Jürgen Leitner, and Peter Corke. DGBench: An Open-Source, Reproducible Benchmark for Dynamic Grasping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3218–3224, Kyoto, Japan, October 2022. IEEE. ISBN 978-1-66547-927-1. doi: 10.1109/IROS47612.2022.9981670. URL https://ieeexplore.ieee.org/document/9981670/.

[3] Alex B. Carter, Catherine Collier, Emma Lawrence, Michael A. Rasheed, Barbara J. Robson, and Rob Coles. A spatial analysis of seagrass habitat and community diversity in the Great Barrier Reef World Heritage Area. *Scientific Reports*, 11(1):22344, November 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-01471-4. URL https://www.nature.com/articles/s41598-021-01471-4.

[4] Haonan Chang, Kowndinya Boyalakuntla, Shiyang Lu, Siwei Cai, Eric Jing, Shreesh Keskar, Shijie Geng, Adeeb Abbas, Lifeng Zhou, Kostas Bekris, et al. Context-aware entity grounding with open-vocabulary 3d scene graphs. *arXiv preprint arXiv:2309.15940*, 2023.

[5] Shengkang Chen, Matthew J. O'Brien, Fletcher Talbot, Jason Williams, Brendan Tidd, Alex Pitt, and Ronald C. Arkin. Multi-modal user interface for multi-robot control in underground environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[6] Nicos Christofides. Worst-case analysis of a new heuristic for the travelling salesman problem. 1976.

[7] Sabit Ekin. Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices. *Authorea Preprints*, 2023.

[8] L. Erhan, M. Ndubuaku, M. Di Mauro, W. Song, M. Chen, G. Fortino, O. Bagdasar, and A. Liotta. Smart anomaly detection in sensor systems: A multi-perspective review. *Information Fusion*, 67:64–79, 2021. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2020.10.001. URL https://www.sciencedirect.com/science/article/pii/S1566253520303717.

[9] Szymon Fedor and Martin Collier. On the problem of energy efficiency of multi-hop vs one-hop routing in wireless sensor networks. In *21st International*

*Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*, volume 2, pages 380–385, 2007. doi: 10.1109/AINAW.2007.272.

[10] Seth G Fitzgerald, Gary W Delaney, and David Howard. A review of jamming actuation in soft robotics. In *Actuators*, volume 9, page 104. MDPI, 2020.

[11] Terrence Fong, Maria Bualat, Laurence Edwards, Lorenzo Flückiger, Clayton Kunz, Susan Lee, Eric Park, Vinh To, Hans Utz, Nir Ackner, et al. Human-robot site survey and sampling for space exploration. In *Space 2006*, page 7425. 2006.

[12] Jesse Haviland, Niko Sünderhauf, and Peter Corke. A holistic approach to reactive mobile manipulation. *IEEE Robotics and Automation Letters*, 7(2):3122–3129, 2022.

[13] Nicolas Hudson, Fletcher Talbot, Mark Cox, Jason Williams, Thomas Hines, Alex Pitt, Brett Wood, Dennis Frousheger, Katrina Lo Surdo, Thomas Molnar, Ryan Steindl, Matt Wildie, Inkyu Sa, Navinda Kottege, Kazys Stepanas, Emili Hernandez, Gavin Catt, William Docherty, Brendan Tidd, Benjamin Tam, Simon Murrell, Mitchell Bessell, Lauren Hanson, Lachlan Tychsen-Smith, Hajime Suzuki, Leslie Overs, Farid Kendoul, Glenn Wagner, Duncan Palmer, Peter Milani, Matthew O'Brien, Shu Jiang, Shengkang Chen, and Ronald Arkin. Heterogeneous ground and air platforms, homogeneous sensing: Team csiro data61's approach to the darpa subterranean challenge. *Field Robotics*, 2(1):595–636, March 2022. ISSN 2771-3989. doi: 10.55417/fr.2022021. URL http://dx.doi.org/10.55417/fr.2022021.

[14] Josie Hughes, Utku Culha, Fabio Giardina, Fabian Guenther, Andre Rosendo, and Fumiya Iida. Soft manipulators and grippers: A review. *Frontiers in Robotics and AI*, 3: 69, 2016.

[15] Jose M Hurtado Jr, Kelsey Young, Jacob E Bleacher, W Brent Garry, and James W Rice Jr. Field geologic observation and sample collection strategies for planetary surface exploration: Insights from the 2010 desert rats geologist crewmembers. *Acta Astronautica*, 90(2):344–355, 2013.

[16] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023.

[17] Navinda Kottege, Jason Williams, Brendan Tidd, Fletcher Talbot, Ryan Steindl, Mark Cox, Dennis Frousheger, Thomas Hines, Alex Pitt, Benjamin Tam, Brett Wood, Lauren Hanson, Katrina Lo Surdo, Thomas Molnar, Matt Wildie, Kazys Stepanas, Gavin Catt, Lachlan Tychsen-Smith, Dean Penfold, Leslie Overs, Milad Ramezani, Kasra Khosoussi, Farid Kendoul, Glenn Wagner, Duncan Palmer, Jack Manderson, Corey Medek, Matthew O'Brien, Shengkang Chen, and Ronald C. Arkin. Heterogeneous robot teams with unified perception and autonomy: How team csiro data61 tied for the top score at the darpa subterranean challenge, 2023.

[18] Weicheng Kuo, Fred Bertsch, Wei Li, AJ Piergiovanni, Mohammad Saffar, and Anelia Angelova. Findit: Generalized localization with natural language queries. In *European Conference on Computer Vision*, pages 502–520. Springer, 2022.

[19] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[20] Lois Liow, James Brett, Josh Pinskier, Lauren Hanson, Louis Tidswell, Navinda Kottege, and David Howard. A compliant robotic leg based on fibre jamming, 2023.

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[22] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[23] Yosio Nakamura, Gary V. Latham, and H. James Dorman. Apollo Lunar Seismic Experiment—Final summary. *Journal of Geophysical Research: Solid Earth*, 87 (S01), November 1982. ISSN 0148-0227. doi: 10.1029/JB087iS01p0A117. URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/JB087iS01p0A117.

[24] Clive R Neal, Charles K Shearer, Meenakshi Wadwha, Lars Borg, Bradley Jolliff, and Allan Treiman. Developing sample return technology using the earth's moon as a testing ground. *Inner Planets Panel, NRC Decadal Survey*, 2013.

[25] Matthew O'Brien, Jason Williams, Shengkang Chen, Alex Pitt, Ronald Arkin, and Navinda Kottege. Dynamic task allocation approaches for coordinated exploration of subterranean environments. *Autonomous Robots*, pages 1–19, 2023.

[26] Lorenzo Pagliara, Vincenzo Petrone, Enrico Ferrentino, and Pasquale Chiacchio. Human-robot interface for teleoperated robotized planetary sample collection and assembly. In *2023 IEEE 10th International Workshop on Metrology for AeroSpace (MetroAeroSpace)*, pages 171–176, 2023. doi: 10.1109/MetroAeroSpace57412.2023.10189984.

[27] Joshua Pinskier and David Howard. From bioinspiration to computer generation: Developments in autonomous soft robot design. *Advanced Intelligent Systems*, 4(1): 2100086, 2022.

[28] Joshua Pinskier, James Brett, Lauren Hanson, Katrina Lo Surdo, and David Howard. Jammkle: Fibre jamming 3D printed multi-material tendons and their application in a robotic ankle. In *IEEE International Conference on Intelligent Robots and Systems*, volume 2022-October, pages 8507–8514. Institute of Electrical and Electronics Engineers Inc., 2022. ISBN 9781665479271. doi: 10.1109/IROS47612.2022.9982171.

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

[30] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=wMpOMO0Ss7a.

[31] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[32] Nicole Robinson, Jason Williams, Gerard Howard, Brendan Tidd, Fletcher Talbot, Brett Wood, Alex Pitt, Navinda Kottege, and Dana Kulic. Human-robot team performance compared to full robot autonomy in 16 real-world search and rescue missions: Adaptation of the darpa subterranean challenge. *arXiv preprint arXiv:2212.05626*, 2022.

[33] Daniela Rus and Michael T Tolley. Design, fabrication and control of soft robots. *Nature*, 521(7553):467–475, 2015.

[34] James F. Russell, D. Klaus, and T. Mosher. Applying analysis of international space station crew-time utilization to mission design. *Journal of Spacecraft and Rockets*, 43:130–136, 2006. doi: 10.2514/1.16135.

[35] Dirk Schulze-Makuch, Louis N Irwin, Dirk Schulze-Makuch, and Louis N Irwin. Optimizing space exploration. *Life in the Universe: Expectations and Constraints*, pages 275–286, 2018.

[36] Hashini Senaratne, Alex Pitt, Fletcher Talbot, Peyman Moghadam, Pavan Sikka, David Howard, Jason Williams, Dana Kulić, and Cécile Paris. Measuring situational awareness latency in human-robot teaming experiments. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 2624–2631, 2023. doi: 10.1109/RO-MAN57019.2023.10309377.

[37] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. Chatgpt for robotics: Design principles and model abilities. Technical Report MSR-TR-2023-8, Microsoft, February 2023. URL https://www.microsoft.com/en-us/research/publication/chatgpt-for-robotics-design-principles-and-model-abilities/.

[38] Georg Wittenburg, Norman Dziengel, Christian Wartenburger, and Jochen Schiller. A system for distributed event detection in wireless sensor networks. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, IPSN '10, page 94–104, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589886. doi: 10.1145/1791212.1791225. URL https://doi.org/10.1145/1791212.1791225.

[39] Matheus S Xavier, Charbel D Tawk, Ali Zolfagharian, Joshua Pinskier, David Howard, Taylor Young, Jiewen Lai, Simon M Harrison, Yuen K Yong, Mahdi Bodaghi, et al. Soft pneumatic actuators: A review of design, fabrication, modeling, sensing, control and applications. *IEEE Access*, 10:59442–59485, 2022.

[40] Mingze Xi, Madhawa Perera, Stuart Anderson, and Matt Adcock. Towards situated imaging. In *IEEE International Conference on Artificial Intelligence extended and Virtual Reality*, pages 85–89. IEEE, 2024. doi: 10.1109/AIxVR59861.2024.00019.

[41] Chaoning Zhang, Dongshen Han, sYu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications, 2023.

[42] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023.