



Delft University of Technology

Document Version

Final published version

Licence

CC BY

Citation (APA)

Sethi, N., Lodel, M., Ferranti, L., Babuška, R., & Alonso-Mora, J. (2026). STEM: Semantic target search and exploration using MAVs in cluttered environments. *Autonomous Robots*, 50(3), Article 32. <https://doi.org/10.1007/s10514-026-10247-6>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.



STEM: Semantic target search and exploration using MAVs in cluttered environments

Nikhil Sethi¹ · Max Lodel¹ · Laura Ferranti¹ · Robert Babuška^{1,2} · Javier Alonso-Mora¹

Received: 12 September 2025 / Revised: 29 January 2026 / Accepted: 14 March 2026
© The Author(s) 2026

Abstract

Autonomous target search is crucial for deploying Micro Aerial Vehicles (MAVs) in emergency response and rescue missions. Existing approaches either focus on 2D semantic navigation in structured environments – which is less effective in complex 3D settings, or on robotic exploration in cluttered spaces – which often lacks the semantic reasoning needed for efficient target search. This paper overcomes these limitations by proposing a novel framework that utilizes a semantically-guided viewpoint planner to minimize target search and exploration time in unstructured 3D environments using an MAV. Specifically, we develop a combinatorial planner that generates efficient semantic exploration plans by prioritizing viewpoints that likely lead to the target. To guide the planner towards the target, an active perception pipeline is developed that propagates semantic priorities of observed objects into neighboring frontier voxels for computing semantic information gains of frontier viewpoints. In addition, we demonstrate how LLM-based similarity scores can be leveraged as semantic priority input to our pipeline. Evaluations in two distinct simulation environments show that the proposed method consistently outperforms baselines by quickly finding the target while maintaining reasonable exploration times. Real-world experiments with an MAV further demonstrate the method’s ability to handle practical constraints like limited battery life, small sensor range, and semantic uncertainty.

Keywords Informative path planning · Search and rescue · Drone exploration · Semantic exploration

1 Introduction

Micro Aerial Vehicles (MAVs) are a promising tool to effectively search and explore complex, unknown environments in domains such as search and rescue, inspection, and envi-

ronmental monitoring. To relieve human operators from the challenging task of guiding MAVs through uncertain environments, methods for autonomous search of target objects are critical to improve the efficiency and effectiveness of such MAV missions.

In large and cluttered environments with many occlusions, the time efficiency of the search strategy is crucial due to limited available flight time and onboard sensor capabilities. Humans can efficiently search unknown spaces by leveraging their experience as well as semantic information, such as observed objects, to reason about the target’s likely location. For example, dangerous chemicals are more likely to be found in a storage room than in an office. Building on this idea, recent work (Chaplot et al., 2020; Chen et al., 2023; Gadre et al., 2022; Ginting et al., 2024; Hahn et al., 2021; Huang et al., 2022; Kim et al., 2022; Ramakrishnan et al., 2022; Yokoyama et al., 2024) has shown that learning such semantic priors can significantly reduce the target search time by guiding the robot toward promising regions. However, as semantic observations or priors can be uncertain or unavailable in many real-world scenarios, purely relying on such

N. Sethi, M. Lodel have contributed equally to this work.

✉ Max Lodel
max.lodel@gmail.com

Nikhil Sethi
sethi.nirvil@gmail.com

Laura Ferranti
l.ferranti@tudelft.nl

Robert Babuška
r.babuska@tudelft.nl

Javier Alonso-Mora
j.alonsomora@tudelft.nl

¹ Department of Cognitive Robotics, Delft University of Technology, Delft, The Netherlands

² CIIRC, Czech Technical University in Prague, Prague, Czech Republic

semantic guidance may lead to inefficient behavior. Therefore, a robust target search strategy should balance semantic search and coverage-maximizing exploration to ensure an efficient and successful search.

Existing works on semantically-guided target search (Chaplot et al., 2020; Chen et al., 2023; Gadre et al., 2022; Ginting et al., 2024; Hahn et al., 2021; Huang et al., 2022; Kim et al., 2022; Ramakrishnan et al., 2022; Yokoyama et al., 2024) have focused on ground robots moving in 2D, while MAVs' 3D capabilities remain underexplored. In particular, MAVs can overcome occlusions in cluttered environments by changing their altitude and, therefore, improve search efficiency. Moreover, these target search methods rely on inferring semantic relationships from large pre-trained models (Yokoyama et al., 2024), which might require fine-tuning for highly specific scenarios. This underlines the need for integrating and balancing semantic search with efficient coverage-maximizing exploration approaches such as Zhou et al. (2021); Cao et al. (2023); Meng et al. (2017). These methods achieve high coverage efficiency by leveraging long-horizon combinatorial planning techniques, but ignore semantic information that could guide the search towards the target.

In this work, we present STEM, a framework for Semantic Target Search and Exploration for MAVs. By building on recent advances in exploration planning and semantics-driven navigation, our framework enables both semantically-guided and efficient exploration in complex 3D environments.

1.1 Related work

In this section, we discuss existing approaches and how they relate to our work, starting with planning approaches for pure coverage exploration, and then focusing on target search approaches that leverage semantic information.

1.1.1 Exploration planning

The problem of navigating a robot autonomously through an unknown environment to build a complete map from sensor observations has been investigated using a variety of different approaches. The fundamental idea of most approaches is to choose robot actions or plans such that future *viewpoints* efficiently minimize unknown space in the environment. Viewpoints are poses in 3D space from which the robot can observe the environment.

Sampling-based approaches randomly sample viewpoints in free space and evaluate their potential information gain about the map. Rapidly Exploring Random Trees (RRT) have been used to plan a local tree of informative viewpoints (Bircher et al., 2016; Dang et al., 2018; Dharmadhikari et al., 2020; Schmid et al., 2020), with Dang et al. (2018) inte-

grating object search by prioritizing salient objects in the environment. However, the high computational costs limit the planning horizon, leading to greedy and inefficient exploration.

Frontier-based exploration methods focus on observing the boundary between known and unknown space, the *frontiers*, to incrementally reduce the unknown space. While early methods (Yamauchi, 1997) just choose the closest frontier as the next observation target, recent work shows that selecting frontiers to maintain high flight speed improves exploration efficiency (Cieslewski et al., 2017). Similarly to sampling-based approaches, these methods lack long-horizon planning capabilities.

Recent works have shown that combining elements of both sampling-based and frontier-based exploration with planning can improve exploration efficiency. The approaches in Zhou et al. (2021); Cao et al. (2023); Huang et al. (2023); Meng et al. (2017); Zhang et al. (2024) sample viewpoints around different frontiers, and then find a time-optimal global plan that connects these viewpoints using combinatorial planning. Following this approach, Meng et al. (2017); Cao et al. (2023); Zhou et al. (2021) plan over frontier viewpoints by solving a Travelling Salesman Problem (TSP). Specifically, the FUEL framework (Zhou et al., 2021) considers drone dynamics in the TSP cost matrix to achieve efficient and agile exploration. In FAEL (Huang et al., 2023), a version of the minimum latency problem is used for planning that prioritizes frontier viewpoints with high coverage gains. Such approaches can be scaled to large environments using coarse global planning (Cao et al., 2023; Zhang et al., 2024).

The authors of FUEL (Zhou et al., 2021) show that their approach achieves efficient and robust exploration performance on a real-world MAV platform in varying, complex 3D environments, due to an effective integration of global and local planning. As we are interested in 3D semantically-guided exploration with MAVs, we build on the FUEL framework (Zhou et al., 2021) as an exploration baseline, extending it with 3D semantic representations and planning to enable semantic target search. Our target search planner uses an MLP-based formulation similar to FAEL (Huang et al., 2023) and integrates semantic information to guide exploration toward target-relevant objects.

1.1.2 Target search

Autonomously searching for a target object in an unknown environment has been primarily investigated in the domain of indoor structured environments such as apartments, where clear semantic relationships between objects exist (Chaplot et al., 2020; Chen et al., 2023; Gadre et al., 2022; Ginting et al., 2024; Hahn et al., 2021; Huang et al., 2022; Kim et al., 2022; Ramakrishnan et al., 2022; Yokoyama et al., 2024).

These approaches differ in the source of learned semantic priors and the planning strategy used to guide the robot toward the target. In earlier works, domain-specific environment datasets are used for training navigation policies using reinforcement learning (RL) (Chaplot et al., 2020; Kim et al., 2022) and training cost-to-go functions using self-supervised learning (Hahn et al., 2021; Ramakrishnan et al., 2022). Conversely, recent works (Chen et al., 2023; Gadre et al., 2022; Ginting et al., 2024; Huang et al., 2022; Yokoyama et al., 2024) use foundation models such as Vision-Language Models (VLM) (Li et al., 2023; Radford et al., 2021) or Large Language Models (LLM) (Devlin et al., 2018) trained on internet-scale data. The works (Chen et al., 2023; Gadre et al., 2022; Huang et al., 2022; Yokoyama et al., 2024) demonstrate zero-shot VLM/LLM-based target search in indoor environments using embedding similarity scores to choose exploration frontiers. In Yokoyama et al. (2024); Gadre et al. (2022); Huang et al. (2022), the frontier selection is facilitated by propagating similarity scores into 2D (Gadre et al., 2022; Yokoyama et al., 2024) or 3D (Huang et al., 2022) map representations. SEEK (Ginting et al., 2024) proposes to distill semantic knowledge from an LLM into a lightweight model for efficient online inference.

The planning strategies used in most of these works are either based on learned navigation policies (Chaplot et al., 2020; Kim et al., 2022) or greedy frontier selection (Chen et al., 2023; Hahn et al., 2021; Ramakrishnan et al., 2022; Yokoyama et al., 2024). In contrast, Ginting et al. (2024) uses a Bayesian network prediction model and value iteration planning to choose the best region to search.

The main limitation of these methods is that they only consider greedy decision-making in 2D and structured indoor environments. In the domain of 3D aerial navigation, recent works have addressed target search by semantics-aware reconstruction of relevant objects (Papatheodorou et al., 2023) and efficient complete coverage of environments (Luo et al., 2024) to find arbitrary targets. In contrast, we aim to leverage the MAV's capabilities for semantically-guided target search, as in the mentioned 2D methods, but for 3D unstructured environments. While (Papatheodorou et al., 2023) uses semantic classes to determine the required reconstruction resolution of objects, we propose a planning pipeline that prioritizes viewpoints by propagating priorities of semantic classes into 3D space, similar to Huang et al. (2022). Our planning pipeline employs a combinatorial planner, generating efficient tours over frontier viewpoints similar to the TSP-based planner in Luo et al. (2024). However, our weighted MLP formulation prioritizes semantically relevant frontiers, allowing for combining semantic target search and efficient exploration. In addition, we demonstrate the use of LLM-based similarity scores as a possible source of semantic priorities, as proposed by Chen et al. (2023).

1.2 Contribution

The main contribution of this paper is a 3D semantic target search framework centered on a novel semantically-guided viewpoint planner designed to minimize expected search time. The combinatorial planner prioritizes frontier viewpoints that likely lead to the target, as evaluated by an active perception pipeline that efficiently propagates semantic priorities from observed objects into continuous, occlusion-aware information gains of 3D candidate viewpoints. Building on the FUEL exploration framework, our novel combinatorial planner formulation, which leverages both semantic and coverage viewpoint gains, generates semantically-guided exploration behavior. We present proof-of-concept experiments in both simulation and real-world environments using a Micro Aerial Vehicle (MAV) that validate the effectiveness of our approach. Our quantitative results show that our method consistently outperforms exploration-only baselines in two distinct simulation environments, in terms of target search time and success rate.

2 Preliminaries

2.1 Problem formulation

An MAV is tasked with exploring a previously unseen 3D environment represented as a bounded volume $\mathcal{W} \subset \mathbb{R}^3$ to find a target object in minimum time. The MAV's pose in the environment at time instant t is defined as $\mathbf{x}_t \in SO(3)$, and we assume that fast and accurate position and attitude controllers are available, such that the robot can follow a trajectory by tracking pose increments (Mellinger & Kumar, 2011). It has a maximum linear velocity v_{max} , maximum acceleration a_{max} , and maximum yaw rate ω_{max} . The robot is equipped with an RGB-D camera that provides a local observation of the environment. At each time instance t the robot receives a measurement tuple $z_t = (\mathbf{x}_t, \mathcal{I}_c, \mathcal{I}_d)$, where $\mathcal{I}_c$ and $\mathcal{I}_d$ are the RGB and depth images, respectively.

Using the RGB images, the robot can perform semantic segmentation to identify objects in the environment. A set $\mathcal{S}$ of possible objects of interest (OOIs) represented by natural language semantic labels is available for segmentation. Importantly, it is assumed that the objects in $\mathcal{S}$ have semantic relationships defined by a function $F : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^+$, that quantifies how semantically related two objects in $\mathcal{S}$ are. It can be exploited to guide the robot towards a target object $o^* \in \mathcal{S}$. A target is considered found when its relative semantic segmentation area in the robot's field of view crosses a threshold λ_{min} .

Problem Statement: Given a target object o^* , a bounded volume $\mathcal{W}$, and the robot's initial pose $\mathbf{x}_0$, find a collision-free global plan ϱ through $\mathcal{W}$ at each time instant t such

that o^* is discovered in minimum time, using the history of observations $z_{0:t}$ and the semantic relationship function F .

2.2 Background

2.2.1 Frontier exploration planning

The goal of robotic exploration is to efficiently build an occupancy map $\mathcal{M}$ of a bounded volume $\mathcal{W}$ using local range observations such as depth images. This map is a 3D volumetric grid of voxels, with each voxel $m_k \in \mathcal{M}$ storing the probability of occupancy P_k . These probabilities are updated using an inverse camera sensor model, and Bayesian Inference (Han et al., 2019). Building an occupancy map of an unknown environment requires the robot to reduce the unexplored space in $\mathcal{M}$.

Frontier-based exploration is an effective approach for reducing unknown space, which first detects a set $\mathcal{F}$ of *frontiers*, i.e., boundary voxels between known and unknown space. Then, it directs the robot to observe these frontiers efficiently. In this work, we build on the recent frontier-based MAV exploration method FUEL presented in Zhou et al. (2021), which is introduced briefly below.

To maintain efficiency, frontiers are clustered into groups of voxels using a region-growing algorithm, leading to a set of clusters $\mathcal{K}$, each with a minimum size F_{min} . That is, a cluster $K_i \in \mathcal{K}$ with $K_i \subseteq \mathcal{F}$ is only valid if $|K_i| \geq F_{min}$. Around these clusters, a set $\mathcal{V}$ of *viewpoints* is sampled, which are poses in free space that can 'view' the frontiers. Viewpoints are further filtered using a minimum information gain I_{min} , considering only those that view at least I_{min} frontier voxels. Unknown space in V can be reduced by finding the most efficient path through high-quality viewpoints. Recent frontier exploration methods such as FUEL (Zhou et al., 2021) and FAEL (Huang et al., 2023) use combinatorial optimization methods to plan a non-myopic global path through the viewpoints in $\mathcal{V}$.

2.2.2 Semantic relationships

Semantics are labels or categories that humans use to classify objects. Humans use accumulated semantic knowledge to derive relationships between objects of interest when looking for targets. For instance, when searching for a `laptop`, we first look for a `table` as opposed to a `toilet` because the former is more correlated with the target object. These relationships are formalized by the semantic relationship function F that maps a set $\mathcal{S}$ of semantic classes represented by language labels to scalar-valued similarity scores.

Large language models (LLM) or vision-language models like CLIP (Radford et al., 2021) or BERT (Devlin et al., 2018) can use their contextual understanding to infer such relationships. These models use a neural network to first transform

the text input to *vector embeddings*, which are real-valued representations of the text in a high-dimensional feature space. Objects that often occur close to each other often have similar vector embeddings, as they appear in similar contexts in the training data. Therefore, the cosine similarity score can be used to approximate the semantic relationship function F between two labels, obtained by calculating the dot product of their vector embeddings $\mathbf{a}$ and $\mathbf{b}$:

$$F(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \quad (1)$$

Such similarity scores have been employed for target search in Chen et al. (2023); Yokoyama et al. (2024).

3 Methodology

3.1 Overview

To motivate our method, consider how humans search for important objects. We infer target-object relations from the environment context, create a mental map of interesting objects, and then search *near* these OOI to find the target. For instance, we might search near a wardrobe to find clothes or under a table to locate a person during an earthquake (see Figure 1). Our method follows the same approach: To quickly find a target, it is essential to minimize unknown space *near* objects of interest that are conceptually and spatially related to the target object. Therefore, we *prioritize* search in specific regions, unlike coverage-based exploration, which tries to minimize *all* unknown space. Additionally, balancing both tasks is necessary to ensure that the method works even under semantic uncertainty and does not incur significant costs in exploration time. Our pipeline consists of three components.

- **Semantic Priority Masking:** This module processes the RGB image to segment objects, ranks them using a priority function, and compresses the segmentation image into a 2D priority mask (see Section 3.2 and Figure 1).
- **Active Perception:** This module uses the generated priority mask, the depth image, and the drone's state to (a) sample a set of 3D frontier viewpoints in free space, and (b) evaluate the semantic information gain of each viewpoint (Section 3.3).
- **Target Search Planner:** This module solves a combinatorial optimization problem over the 3D viewpoints to create a global plan that prioritizes high-gain viewpoints, minimizing the expected search time (Section 3.4).

Figure 2 visualizes the active perception and planning parts of the pipeline.

3.2 Semantic priority masking

The goal of this module is to use the RGB image $\mathcal{I}_c$ to generate a priority mask image $\mathcal{I}_p$ that has pixel-wise discrete priority values for each object of interest. A semantic segmentation image $\mathcal{I}_s$ and a set $\mathcal{S}_t \in \mathcal{S}$ of currently visible classes are generated using $\mathcal{I}_c$ at time t . The user-defined set $\mathcal{S}$ consists of a variety of relevant objects that can be encountered in the present scenario. In this work, we assume the existence of a learning-based method, such as Mask-RCNN (He et al., 2017) or Fast-SAM (Zhang et al., 2023), that can generate $\mathcal{I}_s$ and $\mathcal{S}_t$.

Instead of using $\mathcal{I}_s$ directly, we compress the semantic segmentation image into a priority mask $\mathcal{I}_p$. This image contains pixel-wise discrete integers for each class, indicating their relative importance to the target class. The process to generate this mask is demonstrated in Figure 1. A priority mapping function $r : \mathcal{S} \rightarrow \mathbb{N}^+$ is derived *offline*, for example, as detailed below, and then queried *online*. The *priority mask* $\mathcal{I}_p$ is created at runtime by replacing each pixel in $\mathcal{I}_s$ with its corresponding priority value.

Offline Priority Inference: In the following, we describe a method to generate a semantic priority function r using offline inference with an LLM. First, situational context about the scenario is added to the target object label in the set $\mathcal{S}$, using the formulation: [label] [preposition] [context], e.g., human in earthquake. This helps to derive more accurate relationships between the target object and other objects in the environment. Second, each class in $\mathcal{S}$ is tokenized and passed through the LLM (bert-large-uncased (Devlin et al., 2018)) to produce an output tensor of embeddings, which is averaged along the sequence dimension, resulting in τ of size $\|\mathcal{S}\| \times n_e$. Here, each row represents a class in the embedding space. The target embedding vector is τ^* . To obtain cosine similarity scores for each class, each vector in τ is compared to τ^* using Eq. (1), producing values between 0 and 1. Finally, the similarity scores are scaled to integer values within the range $[1, p_{\max}]$. Here, the maximum priority value $p_{\max}$ is a tunable parameter controlling the sensitivity of the semantic search. The resulting priority mapping function r from the set of classes to integer-valued priorities is stored offline, and then queried online with the set $\mathcal{S}_t$.

3.3 Active perception

The goal of the active perception module is to use the priority mask $\mathcal{I}_p$, the depth image $\mathcal{I}_d$, and robot pose $\mathbf{x}_t$ to create a set of viewpoints $\mathcal{V}$ in free space and corresponding information gains I for each viewpoint.

Section 3.3.1 provides a method for mapping of semantic priorities in 3D and Section 3.3.2 describes a method to diffuse semantic priorities to neighboring frontiers. Section

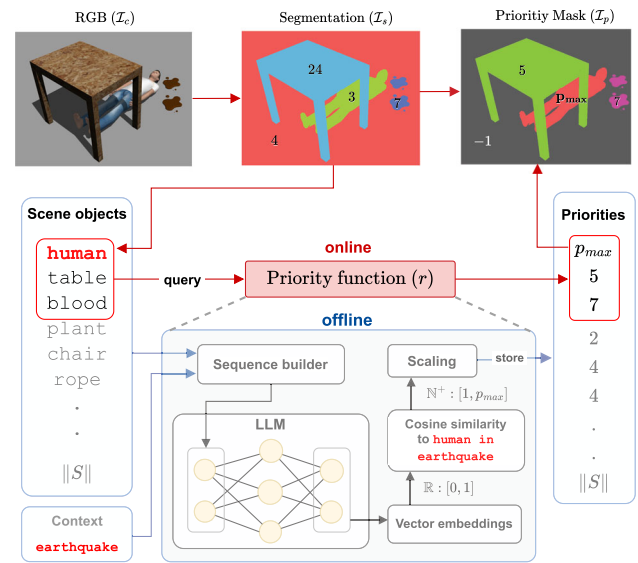


Fig. 1 Semantic priority masking with LLM-based priority function. Red arrows and blue arrows represent online and offline operations, respectively. At runtime, the priority of each class in $\mathcal{S}_t$ is queried from a pre-computed priority vector to create the priority mask $\mathcal{I}_p$

3.3.3 describes the process of generating viewpoints in free space, and Section 3.3.4 describes our procedure to calculate information gain for each sampled viewpoint. Figure 2 shows an overview of the complete active perception pipeline.

3.3.1 Priority map

The goal of this module is to represent the priority values in $\mathcal{I}_p$ in 3D space, where the drone will subsequently navigate and collect new observations. The priority mask $\mathcal{I}_p$, the depth image $\mathcal{I}_d$, and robot pose $\mathbf{x}_t$ are used to create a 4D position-intensity point cloud observation Ω . Each point $\Omega_k = [\mathbf{q}_w, p_w]^T$ in this point cloud is represented by the 3D position in the world frame $\mathbf{q}_w = [x_w, y_w, z_w]^T$ and the priority value p_w as the intensity channel. Let $\mathbf{d}_i = [u_i, v_i]^T$ be a pixel in the depth image $\mathcal{I}_d$ with depth value $z_i = \mathcal{I}_d(\mathbf{d}_i)$ and $p_w = \mathcal{I}_p(\mathbf{d}_i)$ be the corresponding priority value. The point cloud Ω is generated using standard projective transformations using the camera's intrinsic and extrinsic parameters.

The point cloud is also post-processed using voxel-grid filtering and statistical outlier removal. The priority value p_w at each 3D point k in Ω is then used to update a discrete volumetric grid $\mathcal{P}$ at the corresponding voxel $p_k \in \mathcal{P}$ using a simple weighted update (Eq. 2)

$$p_k \leftarrow (1 - \alpha)p_k + \alpha p_w \quad \forall k \in \{1, \dots, |\Omega|\}. \quad (2)$$

Here α is a learning rate that updates the map progressively and prevents noise from being integrated, and $|\Omega|$ is the size of the point cloud. Finally, a local section $\mathcal{P}_l \subset \mathcal{P}$ of the

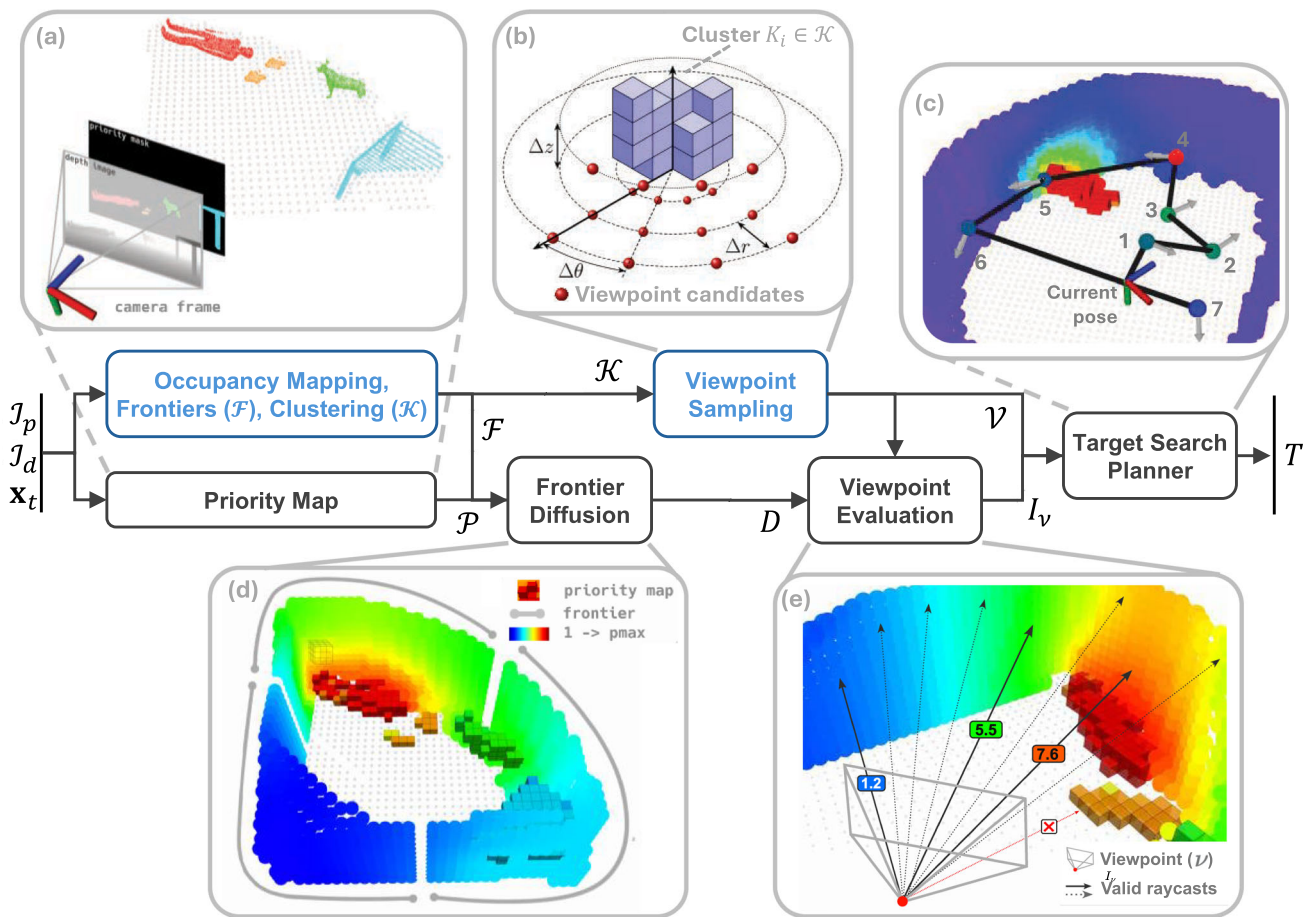


Fig. 2 Active perception and planning pipeline. The blue blocks are from the FUEL framework (Zhou et al., 2021), and Subfigure (b) is inspired by Zhou et al. (2021). In (d), the priority map voxels are cubes, and frontier voxels are spheres

priority map centered around the drone is retrieved to keep the computational load bounded.

Representing semantics as a single-channel scalar priority avoids the computational complexity of high-dimensional 3D semantic mapping or instance tracking. This reduces the mapping task to a simple problem computationally comparable to standard occupancy mapping.

3.3.2 Frontier diffusion

The motivation for this section directly draws from the goal of minimizing unknown space *near* objects of interest (Section 3.1). If frontiers can have increased importance near objects of interest, we can refine the search to interesting regions of the environment and find the target faster.

This idea is implemented by diffusing the priority values from the local map section $\mathcal{P}_l$ into neighboring frontier voxels using a 3D partial convolution. We use a partial convolution for the diffusion process as it normalizes over only the valid (non-empty) voxels in the sparse frontier structure (Liu et al., 2018). A Gaussian kernel with spread σ , and size W is

used, thus making it a 3D Gaussian filter. Figure 2 shows a simulation example from RViz, where the diffusion process is applied to 3D frontiers. The diffusion process is applied to each frontier voxel in a *local* region surrounding the drone to maintain computational efficiency.

Diffusion transforms discrete priority values into a continuous spatial utility function, preventing sparsely sampled viewpoints from missing important frontiers and being discarded by the planner (Section 3.4). Furthermore, diffusing priorities into the volumetric map enables semantic viewpoint evaluations to account for 3D occlusions via raycasting (Section 3.3.4), ensuring the planner prioritizes semantically relevant frontiers that are visually accessible.

The frontier diffusion module thus results in semantic priorities attached to each frontier voxel, that is, a frontier priority function $D : \mathcal{F} \rightarrow [1, p_{\max}]$, which can further be used for downstream tasks like informative path planning.

3.3.3 Viewpoint sampling

The goal of this section is to generate a set of viewpoints $\mathcal{V}$ which are candidate poses sampled in free space to 'view' the frontiers in $\mathcal{F}$. This module directly uses the approach from the FUEL framework (Zhou et al., 2021), which we briefly describe here.

The viewpoint generation is based on the frontier clusters $\mathcal{K}$ introduced in Section 2.2.1. Since even small regions of space can hold significance in semantic target search, we set the minimum cluster size threshold used in FUEL to $F_{\min} = 0$. Frontier viewpoints $v \in \mathcal{V}$ are generated by uniformly sampling free-space poses in a cylindrical coordinate system around the centroid of each frontier cluster (see Figure 2c). The yaw angle of each viewpoint is set to maximize the sensor coverage of the frontier cluster. For more details, see Zhou et al. (2021).

3.3.4 Viewpoint evaluation

This section describes a heuristic for computing a balanced coverage and semantic information gain for a viewpoint $v \in \mathcal{V}$, using the diffusion-based frontier priorities $D(f)$ of each frontier voxel $f \in \mathcal{F}$.

Consider Figure 2d, where the frontier voxels are colored based on their priorities. Rays are cast from a candidate viewpoint v toward the voxels in $\mathcal{F}$ to determine the priority value at the ends of valid rays. A ray is considered valid when it is unobstructed by occupied or unknown space in $\mathcal{M}$. Voxels at the end of valid rays create a new *visible* frontier set $\mathcal{F}_v \subset \mathcal{F}$ for each viewpoint $v \in \mathcal{V}$. The frontier priority value $D(f)$ for each $f \in \mathcal{F}_v$ is then passed through a transfer function Φ and summed up to give the total information gain $I(v)$ of the viewpoint v , as in

$$\Phi(f) = \max \left\{ \exp(\gamma(D(f) - 1)), 1 \right\} \quad (3)$$

$$I(v) = \sum_{f \in \mathcal{F}_v} \Phi(f), \quad (4)$$

where $\gamma \in \mathbb{R}^+$ is a tuning parameter.

To demonstrate this, consider high-priority frontier voxels near semantic objects (as in Figure 2d), which are exponentially weighted by the transfer function Φ . Thus, viewpoints oriented towards semantically meaningful regions of the 3D space achieve a high information gain I_v and can be prioritized in semantic target search. Conversely, consider a frontier f far from semantically interesting objects with the lowest priority value, i.e., $D(f) = 1$, resulting in a volumetric coverage gain $\Phi(f) = 1$. The parameter γ weights the semantic priority against the coverage gain: a higher γ increases the difference between semantically relevant and irrelevant areas, while a lower γ shifts the objective towards

coverage exploration. For $\gamma = 0$, the viewpoint gain equals the number of covered voxels, i.e., $I(v) = |\mathcal{F}_v|$.

The described method for information gain evaluation can integrate semantic priorities with volumetric coverage and uses a tunable exponential weighting function to ensure that semantically relevant viewpoints are prioritized.

3.4 Target search planner

The goal of the global target search planner is to use the set of viewpoints $\mathcal{V}$ and their respective information gains I generated in the active perception module together with the drone's state $\mathbf{x}_t$ to plan a global path that minimizes the time to find the target object.

Our approach extends the idea of combinatorial planning between different viewpoints (Huang et al., 2023; Zhou et al., 2021) to determine their optimal visitation order. While FUEL (Zhou et al., 2021) uses a TSP that minimizes the total traveling distance, semantic target search needs to prioritize viewpoints with high semantic information gain, which are expected to be close to the target. To this end, we propose to formulate the combinatorial target search problem as a variant of the Minimum Latency Problem (MLP) (Blum et al., 1994) that minimizes the average waiting time, or *latency*, of multiple tasks, which in our case are the viewpoints. Specifically, our *weighted* MLP (WMLP) formulation prioritizes minimizing the latency of semantically promising viewpoints, using the information gains $I(v)$ as weights for each viewpoint $v \in \mathcal{V}$. This planner formulation is similar to our concurrent work (Lodel et al., 2026), where we focus on leveraging priorities learned from human inputs, while this paper emphasizes efficient target search with an MAV.

The WMLP is formulated over a modified set of viewpoints $\mathcal{V}^*$: Firstly, we only consider viewpoints with a minimum information gain $I_{\min}$, i.e., $\mathcal{V}' = \{v \in \mathcal{V} \mid I(v) \geq I_{\min}\}$, similar as done in Zhou et al. (2021). Secondly, we add the drone's current pose $\mathbf{x}_t$, i.e., $\mathcal{V}^* = \mathcal{V}' \cup \{\mathbf{x}_t\}$, planning over $N = |\mathcal{V}^*|$ poses. The tour T describes a visitation order of the viewpoints $v \in \mathcal{V}^*$, where $T(i) = v_j$ denotes the i^{th} viewpoint in the tour being v_j , for $i, j \in \{0, \dots, N-1\}$. By definition, $T(0) = \mathbf{x}_t$, as each tour starts at the drone's current pose. Let $\mathbf{C} \in \mathbb{R}^{N \times N}$ be a cost matrix quantifying the traveling time between viewpoints, where C_{ij} is the time required to move from viewpoint v_i to v_j . Then the WMLP is formulated as

$$\min_T \sum_{i=1}^N I(T(i)) \sum_{j=1}^i C_{T(j-1), T(j)}, \quad (5)$$

where the inner sum computes the latency until visiting the i^{th} viewpoint $T(i)$, and the tour cost function is a weighted sum of these latencies using the viewpoints' information gains

$I(T(i))$. This formulation extends the classical MLP with uniform weights to weighted task latencies.

Solving the problem in Eq. (5) means that viewpoints v with higher information gain $I(v)$ get scheduled earlier in the tour. Since the information gain was calculated by balancing coverage and semantic priorities in Section 3.3.4, this objective pursues both coverage and target search, and is thus more robust to semantic uncertainty.

While defining the cost matrix $\mathbf{C}$ using Euclidean distances is common in TSP formulations and robotic exploration (Cao et al., 2023; Huang et al., 2023), it does not account for the complex dynamics of MAVs. Thus, we use the kinematic cost function from Zhou et al. (2021) to compute the traveling times in the cost matrix $\mathbf{C}$. C_{ij} is then defined as the minimum time required to switch between two viewpoints $v_i, v_j \in \mathcal{V}^*$.

$$C_{ij} = \max \left(\frac{\text{length}(v_i^{\mathbf{p}}, v_j^{\mathbf{p}})}{v_{\max}}, \frac{|v_i^{\psi} - v_j^{\psi}|}{\omega_{\max}} \right) \quad (6)$$

Here, $v_i^{\mathbf{p}}$ is the 3D position, $\text{length}(v_i^{\mathbf{p}}, v_j^{\mathbf{p}})$ computes the path length between $v_i^{\mathbf{p}}$ and $v_j^{\mathbf{p}}$, and v_i^{ψ} is the yaw angle of the i^{th} viewpoint from set $\mathcal{V}^*$.

The objective in Eq. (5) is approximately solved using a large neighborhood search (LNS) algorithm (Pisinger & Ropke, 2019) that iteratively destroys and reconstructs the tour T . In each iteration, the custom LNS algorithm randomly removes up to 30% of the tour and then reconstructs it using the cheapest insertion heuristic (Rosenkrantz et al., 1977) followed by a 2-opt swapping search (Croes, 1958).

In summary, the tour T minimizes the time to arrive at regions of the environment that are semantically important with respect to the target object, thus approximately minimizing target search time.

4 Experimental setup

4.1 Simulation environments

Two realistic simulation environments were used to evaluate the algorithm in the PX4-Gazebo SITL simulator¹. The *Earthquake* is a custom environment (Figure 3), and the *Cave* (Figure 4) is a section of the 'Cave Circuit 02' world from the DARPA SubT Challenge².

Both environments contain a trapped human as the target, sufficiently occluded to make the problem challenging. A common superset of objects was used as semantic clues for both environments. This set combines object classes from

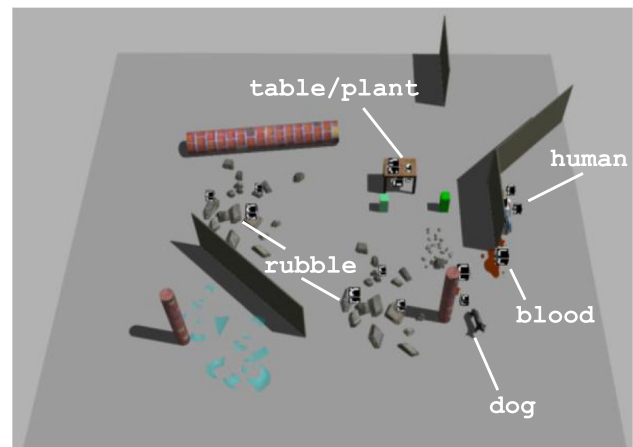


Fig. 3 Earthquake environment (Gazebo)

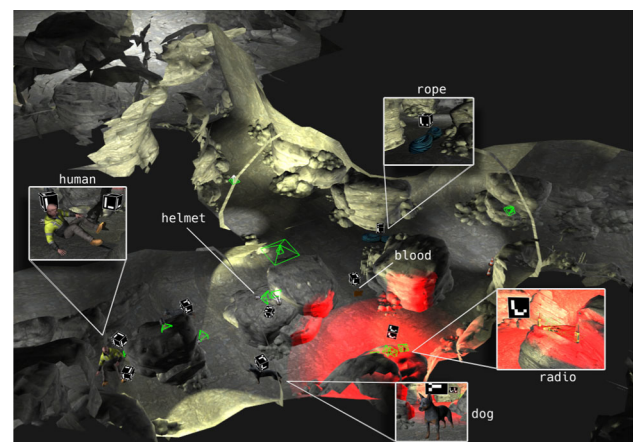


Fig. 4 Cave environment (Gazebo)

the DARPA SubT challenge and common sense objects that are expected near a human target in search and rescue situations. These objects are placed in the environment according to their semantic relationship with the target, e.g., blood is the closest to the human, while the rubble is further away. We also included unimportant distractor objects like toy and plant to evaluate whether the planner prioritizes semantically relevant objects rather than any observed object.

We used ArUco markers as a robust proxy for 2D semantic segmentation. This decouples the evaluation of our target search pipeline from the performance fluctuations of specific segmentation models, ensuring that the results reflect the efficiency of the search strategy rather than the robustness of the vision module. These markers were placed near their respective semantic objects, and the 2D segmentation image then contains a pixel-wise label for each marker $o \in \mathcal{S}$.

Both environments were made sufficiently large to ensure realistic exploration, and the start pose was chosen to be far from the target to observe the effect of semantic exploration. Since we evaluate the experiments on the absolute time-to-

¹ docs.px4.io/main/en/simulation/ros_interface.html

² <https://www.darpa.mil/program/darpa-subterranean-challenge>

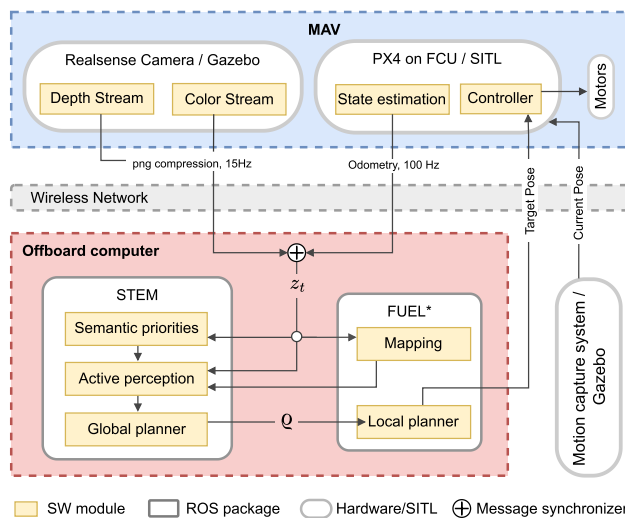


Fig. 5 Software architecture. A modified version of FUEL (Zhou et al., 2021) is used for mapping and local planning

target metric rather than a relative metric (see Section 5.1), starting from random start poses in a relatively small environment will not give comparable results. However, it was observed that there is substantial variance in the SITL simulation (Section 4.2), primarily caused by non-deterministic communication between different software nodes, resulting in different observed trajectories for multiple runs with the same start pose. Therefore, we ran the same experiment multiple times for quantitative evaluation in each environment instead of varying the start pose. In the simulation experiments, we use the offline LLM inference described in Section 3.2 as the source of the semantic priority function r .

4.2 Software architecture

Our software stack is based on the Robotics Operating System (ROS) and integrates our target search method described in Section 3 with the modified FUEL exploration pipeline (Zhou et al., 2021). The architecture is demonstrated in Figure 5. The FUEL pipeline contributes the implementations for 3D voxel mapping, frontier detection and clustering, and viewpoint sampling (see also Figure 2), as well as the local planner. A key capability of the software is that we use the same pipeline for both hardware and software experiments, with the only difference being the source of the measurement tuple z_t . For simulation, this measurement comes from the Gazebo simulator, whereas for hardware experiments, this measurement comes from the onboard camera and state-estimation based on motion capture and IMU information. The parameters used by the planning pipeline are summarized in Table 1.

Table 1 Parameter values for our pipeline used in all experiments

Parameter	Value	Parameter	Value
α	0.9	v_{max}	0.5 m/s
a_{max}	0.5 m/s^2	ω_{max}	0.7 rad/s
h	480	w	848
σ	2	W	5
p_{max}	8	γ	4
I_{min}	10	λ_{min}	0.01
$\ S\ $	22	F_{min}	0

4.3 Hardware setup

Hardware experiments were performed with a custom-built Micro Aerial Vehicle (MAV) previously used in Cao et al. (2023) and modeled for the Gazebo simulation experiments. The MAV is equipped with an Intel RealSense D455 camera and an Nvidia Jetson Xavier NX onboard computer. A HolyBro Kakute F7 V2 flight controller was used with PX4 autopilot software. The MAV was localized in the environment via a Vicon motion capture system. ArUco markers were placed in the environment as semantic objects of interest, and the experiments were conducted in sufficiently cluttered configurations with screens and boxes as obstacles. The lab environments are created to be geometrically challenging due to obstacles and occlusions, but do not represent real scenarios with actual semantic relationships. Therefore, in the hardware experiments, we do not use LLM-based priority inference (see Section 3.3.1) used in the simulation experiments, but instead provide a handcrafted priority function r . Note that in this case, the whole pipeline, including priority masking, active perception, and planning, is still validated in the hardware experiments. Figure 6 shows an exemplary environment for the experiment.

5 Simulation results

In this section, we present and discuss the results of our proposed method *STEM* on the simulation environments from Section 4.1. Evaluation metrics are proposed in Section 5.1, and two baselines are compared in Section 5.2. We provide the primary performance comparisons with baselines in Section 5.3, which are further supported using qualitative results in Section 5.4. Finally, we also conduct an ablation study to analyze the importance of the proposed target search planner in Section 5.5.

5.1 Evaluation metrics

We employ the following metrics:

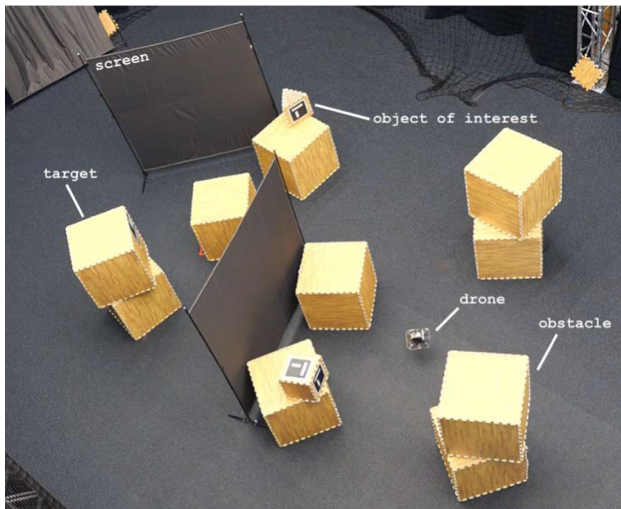


Fig. 6 Lab environment for hardware experiments

Success %: This metric calculates the percentage of episodes in which the target was successfully found across n trials. An object o is considered found when the fraction of object pixels λ_o crosses a threshold λ_{min} in the segmentation image, with λ_o defined as $\lambda_o = \beta / (wh)$. Here, β is the number of pixels belonging to the object's ArUco marker, and w and h are the segmentation image width and height, respectively. The parameter λ_{min} depends on the environment complexity and camera intrinsics. For example, a value of $\lambda_o = 0.02$ means that the marker occupies 2% of the field of view in the image plane.

Time to target: A commonly used metric for ObjectNav tasks is Success weighted by Path Length (SPL) (Batra et al., 2020). A notable drawback of this metric is that it only considers travel distance in SE(3), and for robots with complex dynamics (such as MAVs), the completion time is recommended (Yokoyama et al., 2021). Thus, we record the first time instant when a target o^* was successfully detected (i.e., $\lambda_{o^*} \geq \lambda_{min}$) and call this metric the Time to target t^* .

Exploration time: Since balancing exploration and target search is a secondary goal for our method, we also measure the exploration time t_f in seconds. An environment is considered explored when no *visible* frontier can be found for 10 consecutive iterations. For a frontier to be considered *visible*, it must have (1) at least F_{min} number of clustered voxels, and (2) at least one viewpoint with minimum information gain I_{min} . These conditions are also used by the authors of FUEL (Zhou et al., 2021).

5.2 Baselines

We evaluate our method by comparing it with the highly efficient coverage-based exploration pipeline FUEL (Zhou et al., 2021), since there are no semantically guided target

search methods for MAVs. This allows us to isolate the performance gain achieved by adding semantic guidance to a coverage framework. Our method is compared to two versions of the FUEL pipeline (Zhou et al., 2021), differing by the two parameters F_{min} and I_{min} : The *FUEL-original* baseline uses the original parametrization with $F_{min} = 100$ and $I_{min} = 20$ as proposed in Zhou et al. (2021). For the *FUEL-complete* baseline, these parameters were tuned to maximize target search success, as we noticed that finding the target in small regions depended strongly on these parameters. For the Earthquake scenario, $F_{min} = 0$ and $I_{min} = 5$ were used, and for the Cave scenario, $F_{min} = 0$ and $I_{min} = 0$. Additionally, frontier down-sampling (see (Zhou et al., 2021)) was turned off due to the narrow passages and small frontier sizes in the Cave environment. The sensor range R_{max} for both baselines was kept the same as our method to make comparisons fair.

5.3 Performance results

This section presents the performance results of our method compared to the two baselines introduced in Section 5.2 using the metrics from Section 5.1. Table 2 summarises results for both the Earthquake and Cave simulation environments. The data was gathered over 20 trials for each method and environment. The results are discussed below.

STEM consistently finds the target faster than all methods and is as successful as the FUEL-complete baseline, which is tuned for high success rates. Moreover, it keeps the exploration times within reasonable bounds. The fast target search times and high success rates have been achieved by the combination of the semantic viewpoint evaluation and the combined target search planner. The viewpoint evaluation based on the semantic frontier priorities diffused from relevant objects and subsequent filtering of viewpoints using I_{min} leads to a viewpoint set biased towards viewing regions of high likelihood of target presence. This allows our method to find the target consistently. The combinatorial target search planner schedules high semantic gain viewpoints earlier, such that the MAV quickly reaches regions where the target is likely to be found. Conversely, FUEL-complete does not use semantic information to guide search and explores irrelevant regions first, therefore taking more time to find the target while still achieving high success rates.

The performance difference between FUEL-complete and STEM underlines the importance of the viewpoint gain threshold I_{min} : A lower I_{min} retains viewpoints for small frontier clusters in tight spaces, and therefore enabling a high success rate for FUEL-complete. However, this also means that many small but unimportant regions are covered, leading to inefficient and slow target search and exploration, as reflected by the exploration time results of FUEL-complete. This emphasizes the advantage of evaluating the semantic relevance of frontiers, as it allows retaining and prioritiz-

Table 2 Comparison study with baselines in the Earthquake and Cave environments. Two baselines from Section 5.2 were compared to our method on success %, time to target t^* , and exploration time t_f

Env	Method	Target search Success %	Time t^* (s)	Exploration Time t_f (s)
Earthquake	FUEL-original	5%	79.8 ± 0.0	107.3 ± 17.3
	FUEL-complete	100%	76.6 ± 33.5	129.4 ± 12.9
	STEM (Ours)	100%	56.9 ± 22.8	139.9 ± 11.9
Cave	FUEL-original	40%	65.0 ± 14.8	93.6 ± 30.7
	FUEL-complete	90%	95.3 ± 26.6	166.4 ± 31.5
	STEM (Ours)	90%	64.1 ± 20.3	130.7 ± 19.8

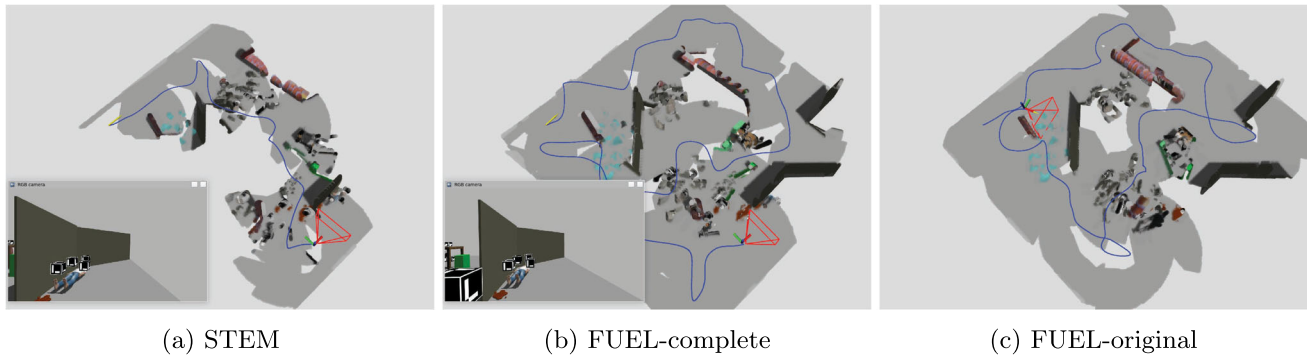


Fig. 7 Qualitative comparison with baselines for the Earthquake environment. Episodes were recorded until t^* or t_f , whichever comes first. The reconstructed point cloud from RViz is shown along with the drone's trajectory (in blue). The camera FOV is shown in red, and the latest RGB camera image is displayed

ing viewpoints in small but important regions while ignoring small, unimportant regions.

FUEL-original rarely finds the target due to its increased minimum frontier size F_{min} and viewpoint threshold I_{min} . As a result, frontier clusters or viewpoints in tight spaces that lead to the target are ignored. However, this baseline consistently completes exploration faster, since the clustering and thresholding facilitate more stable viewpoint sets $\mathcal{V}$ and more consistent global plans, allowing the MAV to maintain high speeds throughout the episode. Furthermore, FUEL-original solves a metric TSP using the LKH heuristic solver, generating robust and efficient global plans.

5.4 Qualitative results

In this section, we visualize the behavior of our proposed method compared to the baselines to support the quantitative results from Section 5.3.

Figure 7 shows a qualitative comparison of the trajectories produced by the three methods in the Earthquake environment until the target is found or exploration is completed. The results show that STEM finds the target quickly without taking large detours, while FUEL-complete explores large parts of the environment before finding the target, and FUEL-original does not find the target. STEM achieves these results by focusing exploration on regions semantically related to the target, while both FUEL methods only aim at maximiz-

ing coverage gains. The trajectory of FUEL-complete shows that it came close to the target and highly relevant semantic cues (blood objects), but does not react to these cues and continues exploring irrelevant regions.

Comparing FUEL-complete and FUEL-original emphasizes the effect of the thresholds F_{min} and I_{min} discussed in Section 5.3. When FUEL-original does not find any sufficiently large frontiers or viewpoints and completes exploration, there are still several small regions of the map left unexplored. FUEL-complete, however, is more thorough in its exploration but explores the environment less efficiently, as evidenced by more directional changes in its trajectory.

Figure 8 further shows keyframes of our algorithm performing target search in the earthquake environment. The MAV starts at a disadvantaged position and starts to explore first to gather information. When it comes across semantically relevant objects such as rubble or blood, it samples informative viewpoints covering the nearby frontiers. Planning a path through such viewpoints, continually using the target search planner, allows the MAV to find the hidden target quickly. In addition, the keyframes show that our approach correctly ignores the distractor objects like "plant".

Figure 9 presents the trajectories of our method and the two baselines in the cave environment. The results show that STEM is able to handle a complex 3D environment with many occlusions and tight spaces, which underlines its abil-

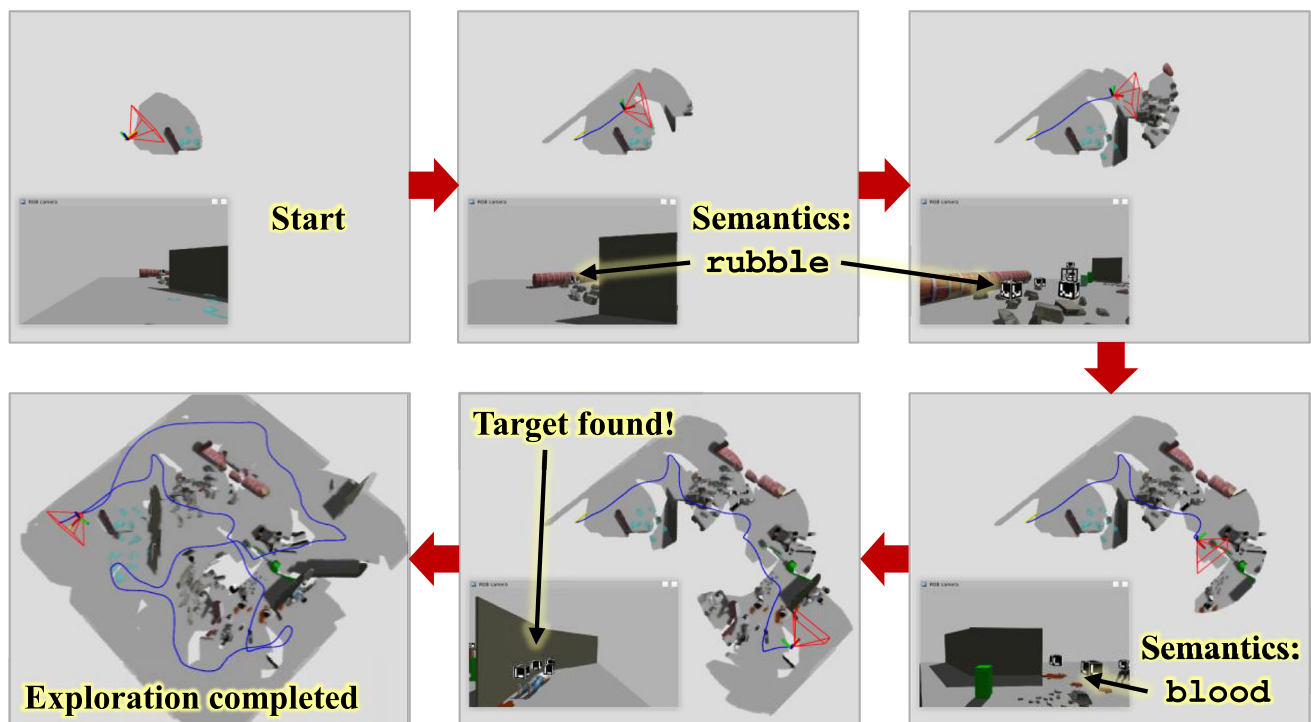


Fig. 8 Key frames of target search using STEM in the Earthquake environment. The snapshots show the recorded RGB point cloud, the camera FOV in red, the trajectory in blue, and the current RGB image

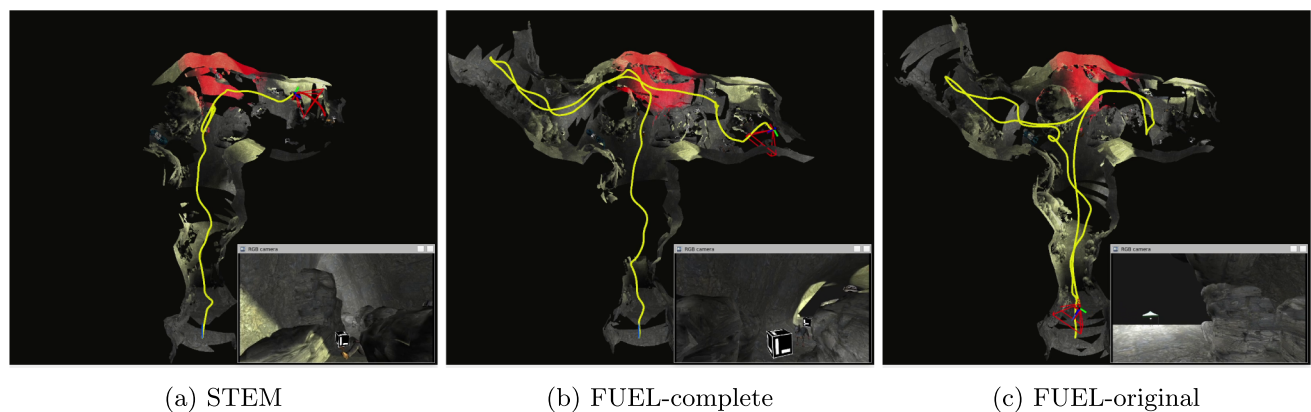


Fig. 9 Qualitative comparison with baselines for the cave environment. Episodes were recorded until t^* or t_f , whichever comes first. The reconstructed point cloud from RViz is shown along with the drone's trajectory (in yellow). The camera FOV is shown in red, and the latest RGB camera image is displayed

ity to balance exploration and semantic target search. After exploring initially, the MAV comes across objects such as radio and dog in the right arm of the cave, which STEM uses to guide the robot to the target quickly. Conversely, FUEL-complete misses the semantic cues and first explores the left arm of the cave before eventually discovering the target. FUEL-original coincidentally explores the right arm of the cave first, but does not find the target due to its thresholding of frontier clusters and viewpoint gains.

5.5 Planner ablation study

In this section, we evaluate the importance of the combinatorial target search planner in our method. We compare our full method with the WMLP-based planner from Section 3.4 with two simplified variants that use the same viewpoint sampling, evaluation, and filtering methods using semantics, but replace the planner. The first variant uses a greedy viewpoint choice, planning to the viewpoint with the highest information gain I_v . The second variant uses the TSP planner from

Table 3 Planner ablation study, comparing our proposed WMLP planner with a greedy and a TSP-based planner on success %, time to target t^* , and exploration time t_f

Env	Method	Target search Success %	Time, t^* (s)	Exploration Time, t_f (s)
Earthquake	Greedy	100%	71.1 ± 39.1	210.5 ± 25.8
	TSP-LKH	100%	65.3 ± 26.7	128.5 ± 10.8
	WMLP (STEM)	100%	56.9 ± 22.8	139.9 ± 11.9
Cave	Greedy	50%	64.9 ± 18.1	175.8 ± 35.8
	TSP-LKH	85%	93.7 ± 19.9	134.1 ± 23.5
	WMLP (STEM)	90%	64.1 ± 20.3	130.7 ± 19.8

Zhou et al. (2021) to plan a path through the viewpoints in $\mathcal{V}$, deploying the LKH heuristic to solve the optimization problem. The quantitative results are shown in Table 3.

The results show that STEM with the WMLP planner outperforms both the TSP and greedy planners on target search metrics in both environments. In the earthquake environment, all three variants achieve a 100% success rate, while in the cave environment, the success rate of the greedy planner is substantially lower than that of WMLP and TSP. This indicates that the complex geometry of the cave environment requires more consistent and efficient behavior of the combinatorial planners than the simpler earthquake environment.

The average target search times t^* show significant differences between the planners in both environments. In the earthquake environment, the TSP planner is 15% slower, and the greedy planner is 25% slower in finding the target than the WMLP planner. This underlines that the WMLP planner prioritizing semantically relevant viewpoints contributes substantially to STEM's target search performance. However, it also shows that the performance gap between FUEL and STEM (Section 5.3) is only partially caused by the planning algorithm, and that the semantic diffusion-based viewpoint evaluation and filtering are also critical elements of the target search strategy. In the cave environment, the greedy planner achieves equally fast search times as WMLP *when successful*, while the TSP planner is 46% slower. This indicates that in some cases, maximizing semantic viewpoint gains via the greedy planner helps to find the target quickly. The greedy strategy can work well when the environment constrains the exploration options (e.g., by tunnels) and when unambiguous semantic cues are continually observed. In contrast, the TSP planner treats all viewpoints as equally important and only optimizes for their efficient coverage.

While WMLP and TSP show very similar exploration times, the greedy planner is significantly less efficient in covering the volume. This is due to the greedy planner's tendency to oscillate between viewpoints with high information gain, leading to long detours and inefficient exploration paths.

In summary, this ablation study shows that combinatorial planning contributes to more consistent performance for both target search and exploration, and that both the seman-

tic viewpoint evaluation and the WMLP planner contribute substantially to the target search performance of STEM.

6 Real-world experiments

This section presents the results of real-world experiments conducted using the software architecture from Section 4.2 and the hardware setup from Section 4.3.

The results are visualized in Figures 10 to 12, showing our algorithm performing simultaneous target search and exploration in complex 3D environments set up in the laboratory. Figure 10 shows a scenario where the drone starts from the bottom right corner in the top-down view, and explores the environment searching for the target and relevant objects, represented by ArUco markers. The target is hidden behind a screen and a lower wall in the upper left section of the lab. The onboard camera images in Figure 10 show how the drone first discovers a high-priority object on top of the wall on the right and a low-priority on the left, and therefore prioritizes the unknown area behind the wall on the right due to the diffusion method (see section 3.3.2). In image 2, a high-relevance object is detected behind the wall, creating additional high-priority frontiers around the pillar in the upper left corner, which leads the WMLP planner to prioritize exploring the area behind the wall and screen further. Therefore, the drone is able to discover the target quickly and without large detours in image 3. Following target discovery, the framework guides the drone to further explore the environment efficiently, facilitated by the viewpoint evaluation method combining coverage and semantic information gain (see Section 3.3.4). This scenario underlines how our framework is able to search for a target using semantic clues in a cluttered 3D environment, guiding the drone to fly over obstacles (the wall close to viewpoint 2) to explore important hidden areas behind it.

Figures 11 and 12 show two additional scenarios, where the drone is able to find the target quickly using the ArUco marker objects as semantic clues, as described above. Both experiments show how our framework is able to guide the drone through different cluttered environments, handling

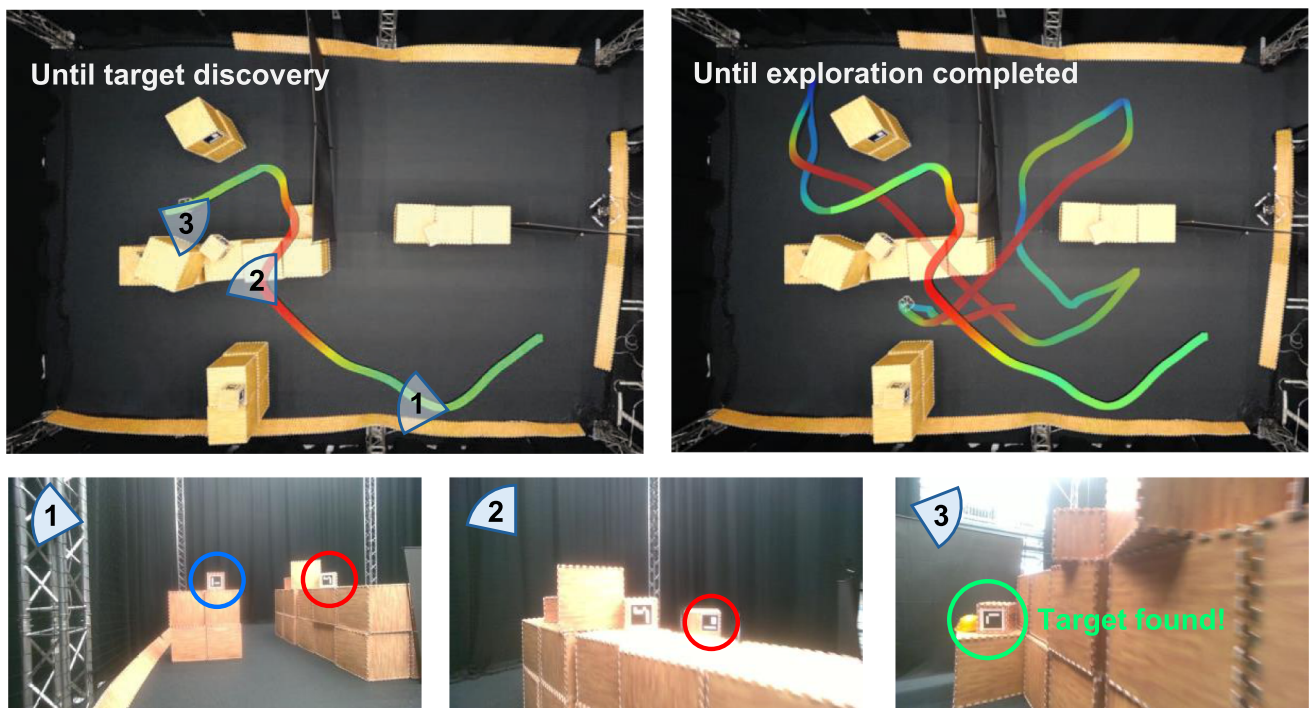


Fig. 10 Hardware experiment 1 in the laboratory. The top row shows two top-down views. The color of the trajectory indicates the drone's altitude z . The bottom row shows images from the onboard RGB camera at three time instances, which are marked by position and orientation in the left top-down view. The red circles in the onboard images indicate

the high-priority objects detected by the framework, which guide the drone towards the target. The blue circle indicates a low-priority distractor object. The last onboard image shows the moment the target is discovered

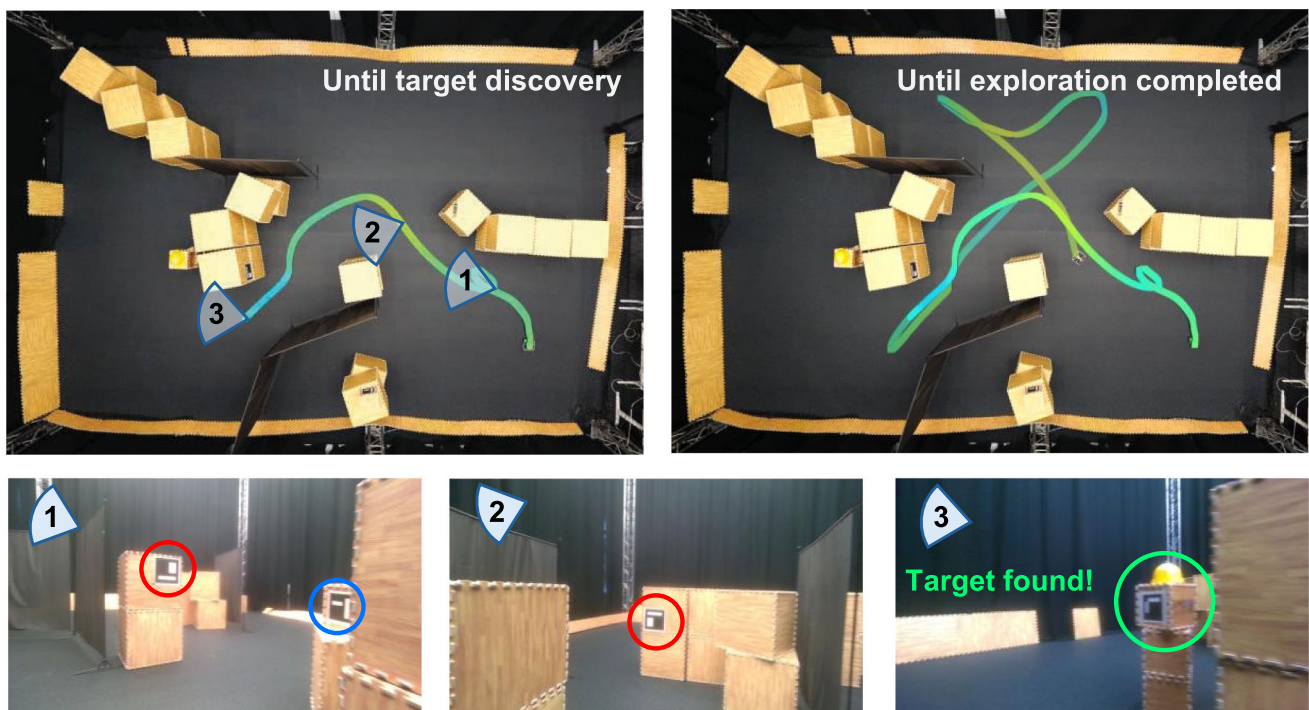


Fig. 11 Hardware experiment 2 in the laboratory, with the same format as Figure 10

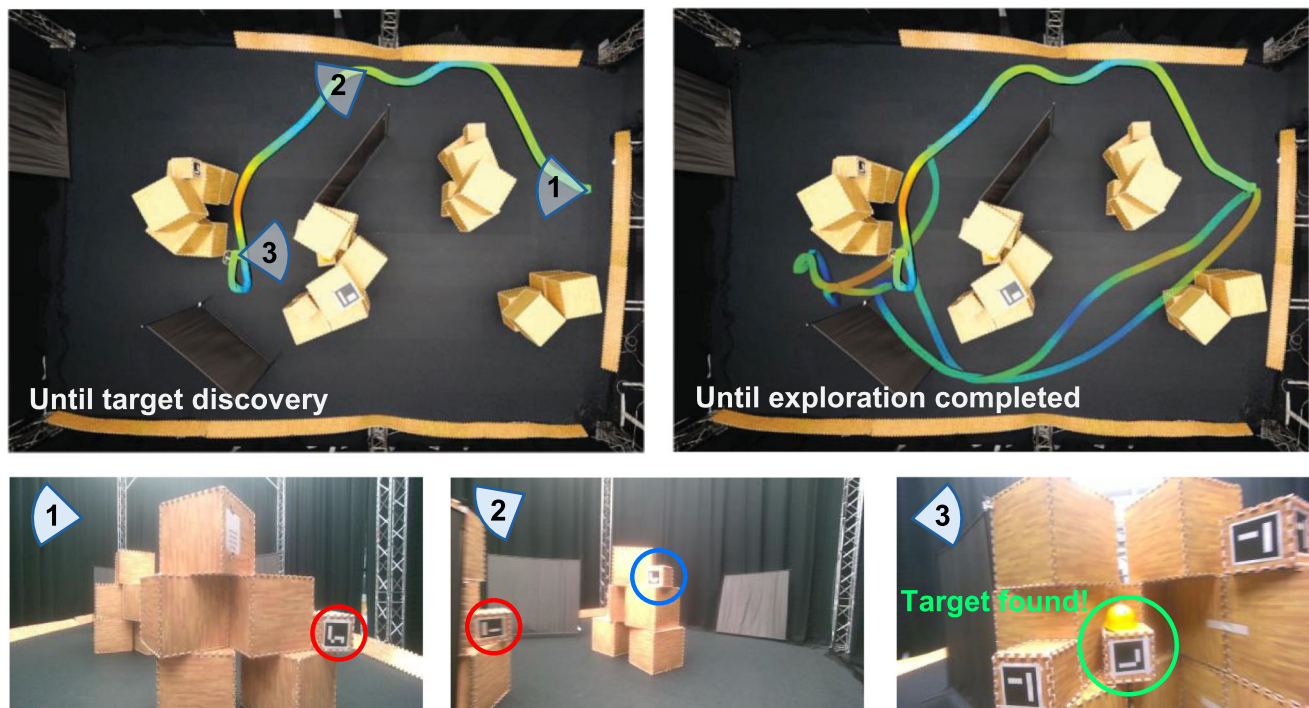


Fig. 12 Hardware experiment 3 in the laboratory, with the same format as Figure 10

occlusions and narrow passages, and to efficiently explore the environment after finding the target.

In summary, the real-world experiments demonstrate the feasibility of our method in handling noisy onboard depth images and real-time constraints, and still efficiently guide the drone towards semantically important areas. This enables the drone to find the target quickly in three different complex 3D environments, and to explore the environment efficiently after target discovery.

7 Conclusion

In this paper, we introduced STEM, a novel framework for semantically-guided target search and exploration with Micro Aerial Vehicles (MAVs) in complex, cluttered 3D environments. The core contribution of this work is the integration of efficient planning techniques commonly used in coverage exploration with semantic relationships guiding target search. The key idea of our approach is to prioritize exploration around semantically relevant objects when planning over all available exploration frontiers. Balancing semantic and coverage objectives in this way, efficient behavior for both target search and exploration is achieved.

Quantitative evaluation in two challenging simulation environments demonstrates that STEM achieves consistently higher success rates of more than 90 % and up to 32% faster target discovery times compared to coverage-maximizing

baseline methods. Real-world experiments further validate the practical applicability of our approach, proving that our approach can exhibit effective target search with noisy camera inputs and realistic drone dynamics. STEM is designed to generalize to diverse environments by balancing semantic and coverage exploration, such that the system defaults to coverage when semantic information is absent or ambiguous. Future work will conduct extensive experiments to evaluate how our method generalizes to environments with diverse topologies and semantic features. It will also investigate how uncertainty about semantic object detection and semantic relationships can be propagated into the active perception pipeline, and how foundation models can be used to infer semantic priorities beyond embedding similarities.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10514-026-10247-6>.

Acknowledgements This work was supported by the National Police of the Netherlands. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Author Contributions M.L. proposed the initial idea of extending frontier-based MAV exploration with semantic priorities, developed and implemented the target search planner, and contributed to the experiments and writing of the paper. N.S. developed and implemented the pipelines for semantic priority masking and active perception and contributed to the experiments and writing of the paper. L.F., R.B. and J.A.-M. provided discussions and feedback on the ideas, methods, results, and the writing of the paper.

Data Availability A video of the drone experiments described in Section 6 is available in the supplementary information of this paper. The code used for simulation and hardware experiments is available here: <https://github.com/tud-amr/stem>.

Declarations

Conflicts of Interest No conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Batra, D., Gokaslan, A., Kembhavi, A., Maksymets, O., Mottaghi, R., Savva, M., Toshev, A. & Wijmans, E. (2020). ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. *arXiv*.
- Bircher, A., Kamel, M., Alexis, K., Oleynikova, H. & Siegwart, R. (2016). Receding Horizon "Next-Best-View" Planner for 3D Exploration. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 1462–1468.
- Blum, A., Chalasani, P., Coppersmith, D., Pulleyblank, B., Raghavan, P. & Sudan, M. (1994). The minimum latency problem. In: *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing - STOC '94*.
- Cao, C., Zhu, H., Ren, Z., Choset, H., & Zhang, J. (2023). Representation granularity enables time-efficient autonomous exploration in large, complex worlds. *Science Robotics*, 8(80), eadf0970. <https://doi.org/10.1126/scirobotics.adf0970>
- Chaplot, D.S., Jiang, H., Gupta, S. & Gupta, A. (2020). Semantic Curiosity for Active Visual Learning. *arXiv*.
- Chen, J., Li, G., Kumar, S., Ghanem, B. & Yu, F. (2023). How To Not Train Your Dragon: Training-free Embodied Object Goal Navigation with Semantic Frontiers. *arXiv*.
- Cieslewski, T., Kaufmann, E. & Scaramuzza, D. (2017). Rapid exploration with multi-rotors: A frontier selection method for high speed flight. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2135–2142.
- Croes, G.A. (1958). A Method for Solving Traveling-Salesman Problems. *Operations Research* 6(6).
- Dang, T., Papachristos, C. & Alexis, K. (2018). Visual Saliency-Aware Receding Horizon Autonomous Exploration with Application to Aerial Robotics. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 2526–2533. IEEE, Brisbane, QLD
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*
- Dharmadhikari, M., Dang, T., Solanka, L., Loje, J., Nguyen, H., Khedekar, N. & Alexis, K. (2020). Motion Primitives-based Path Planning for Fast and Agile Exploration using Aerial Robots. In: 2020 IEEE International Conference on Robotics and Automation (ICRA).
- Gadre, S.Y., Wortsman, M., Ilharco, G., Schmidt, L. & Song, S. (2022). CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation.
- Ginting, M.F., Kim, S.-K., Fan, D.D., Palieri, M. & Kochenderfer, M.J. (2024). Agha-Mohammadi, A.-a.: SEEK: Semantic Reasoning for Object Goal Navigation in Real World Inspection Tasks. *arXiv*.
- Hahn, M., Chaplot, D.S., Tulsiani, S., Mukadam, M., Rehg, J.M. & Gupta, A. (2021). No RL, No Simulation: Learning to Navigate without Navigating. In: *Advances in Neural Information Processing Systems*, vol. 34.
- Han, L., Gao, F., Zhou, B. & Shen, S. (2019). Fiesta: Fast incremental euclidean distance fields for online motion planning of aerial robots. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4423–4430. IEEE
- He, K., Gkioxari, G., Dollár, P. & Girshick, R. (2017). Mask r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969.
- Huang, C., Mees, O., Zeng, A. & Burgard, W. (2022). Visual Language Maps for Robot Navigation.
- Huang, J., Zhou, B., Fan, Z., Zhu, Y., Jie, Y., Li, L., & Cheng, H. (2023). FAEL: Fast Autonomous Exploration for Large-scale Environments With a Mobile Robot. *IEEE Robotics and Automation Letters*, 8(3), 1667–1674.
- Kim, N., Kwon, O., Yoo, H., Choi, Y., Park, J. & Oh, S. (2022). Topological Semantic Graph Memory for Image-Goal Navigation. In: 6th Annual Conference on Robot Learning.
- Li, J., Li, D., Savarese, S. & Hoi, S. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: *Proceedings of the 40th International Conference on Machine Learning*.
- Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A. & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 85–100.
- Luo, Y., Zhuang, Z., Pan, N., Feng, C., Shen, S., Gao, F., Cheng, H. & Zhou, B. (2024). Star-searcher: A complete and efficient aerial system for autonomous target search in complex unknown environments. *IEEE Robotics and Automation Letters*.
- Mellinger, D. & Kumar, V. (2011). Minimum snap trajectory generation and control for quadrotors. In: 2011 IEEE International Conference on Robotics and Automation, pp. 2520–2525. IEEE.
- Meng, Z., Qin, H., Chen, Z., Chen, X., Sun, H., Lin, F., & Ang, M. H. (2017). A Two-Stage Optimized Next-View Planning Framework for 3-D Unknown Environment Exploration, and Structural Reconstruction. *IEEE Robotics and Automation Letters*, 2(3), 1680–1687.
- Papatheodorou, S., Funk, N., Tzoumanikas, D., Choi, C., Xu, B. & Leutenegger, S. (2023). Finding Things in the Unknown: Semantic Object-Centric Exploration with an MAV. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 3339–3345.
- Pisinger, D. & Ropke, S. (2019). Large Neighborhood Search. In: Gendreau, M., Potvin, J.-Y. (eds.) *Handbook of Metaheuristics*.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*, pp. 8748–8763. PMLR.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In: *Proceedings of the 38th International Conference on Machine Learning*

- Ramakrishnan, S.K., Chaplot, D.S., Al-Halah, Z., Malik, J. & Grauman, K. (2022). PONI: Potential Functions for ObjectGoal Navigation with Interaction-free Learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Rosenkrantz, D.J., Stearns, R.E. & Lewis, P.M. (1977). II: An Analysis of Several Heuristics for the Traveling Salesman Problem. *SIAM Journal on Computing* 6(3)
- Schmid, L., Pantic, M., Khanna, R., Ott, L., Siegwart, R., & Nieto, J. (2020). An Efficient Sampling-Based Method for Online Informative Path Planning in Unknown Environments. *IEEE Robotics and Automation Letters*, 5, 1–1.
- Wu, S. (2023). Risk-aware decentralized multi-mav planning in unknown and dynamic environments. MSc. Thesis, Delft University of Technology.
- Yamauchi, B. (1997). A frontier-based approach for autonomous exploration. In: Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97. 'Towards New Computational Principles for Robotics and Automation', pp. 146–151. IEEE
- Yokoyama, N., Ha, S. & Batra, D. (2021). Success weighted by completion time: A dynamics-aware evaluation criteria for embodied navigation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1562–1569. IEEE.
- Yokoyama, N., Ha, S., Batra, D., Wang, J. & Bucher, B. (2024). VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA).
- Zhang, Y., Chen, X., Feng, C., Zhou, B. & Shen, S. (2024). FALCON: Fast Autonomous Aerial Exploration Using Coverage Path Guidance. *IEEE Transactions on Robotics*.
- Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.-H., Lee, S. & Hong, C.S. (2023). Faster segment anything: Towards lightweight sam for mobile applications. [arXiv:2306.14289](https://arxiv.org/abs/2306.14289).
- Zhou, B., Zhang, Y., Chen, X., & Shen, S. (2021). Fuel: Fast UAV exploration using incremental frontier structure and hierarchical planning. *IEEE Robotics and Automation Letters*, 6(2), 779–786.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Nikhil Sethi received his M.Sc. in robotics from Delft University of Technology in 2024. His master's thesis research at the Department of Cognitive Robotics focused on semantic target search and exploration using MAVs.



Max Lodel received his Ph.D. degree in robotics from Delft University of Technology in 2025. His Ph.D. research at the Department of Cognitive Robotics focused on combining learning and planning methods for autonomous search in unknown environments. He received his M.Sc. degree from the Technical University of Freiberg in 2020.



Laura Ferranti received her Ph.D. degree in systems and control from the Delft University of Technology, Delft, The Netherlands, in 2017. She is currently an Associate Professor in the Cognitive Robotics Department at Delft University of Technology, where she leads the Reliable Robot Control Lab. Her research interests include optimization and optimal control, model predictive control, reinforcement learning, and embedded optimization-based control, with applications in mobile robotics.

Dr. Ferranti was the recipient of an NWO Veni Grant from the Netherlands Organization for Scientific Research (2020) and of the Best Paper Award in Multi-robot Systems at ICRA 2019.



Robert Babuska received the M.Sc. (Hons.) degree from the Czech Technical University in Prague in 1990 and the Ph.D. (cum laude) degree from Delft University of Technology, the Netherlands, in 1997. He is a full professor at TU Delft, Faculty of Mechanical Engineering, Department of Cognitive Robotics, and a distinguished researcher at the Czech Technical University in Prague, CIIRC, Department of AI. His research interests include reinforcement learning, adaptive and learning

robot control, and nonlinear system identification. He has been involved in applying these techniques across various fields, ranging from process control to robotics and aerospace.



Javier Alonso-Mora is a full professor at the Cognitive Robotics department of the Delft University of Technology, where he leads the Autonomous Multi-Robots Lab and serves as Director of the TU Delft Robotics Institute. Dr. Alonso-Mora was a post-doctoral associate at the Massachusetts Institute of Technology (MIT). He received his Ph.D. degree in robotics from ETH Zürich. His research focuses on navigation, motion planning, learning, and control of autonomous mobile robots and teams, interact-

ing with humans in dynamic, uncertain environments. He is the recipient of multiple awards, including the IEEE ITS Young Researcher/Engineer Award (2025), the IEEE T-ASE Best Paper Award (2024), an ERC Starting Grant (2021), and the ICRA Best Paper Award on Multi-Robot Systems (2019).