Uncertainty-Informed Active Perception for Open Vocabulary Object Goal Navigation

Utkarsh Bajpai Julius Rückin Cyrill Stachniss Marija Popović

Abstract— **Mobile robots exploring indoor environments increasingly rely on vision-language models to perceive high-level semantic cues in camera images, such as object categories. Such models offer the potential to advance robot behaviour for tasks such as object-goal navigation (ObjectNav), where the robot must locate objects specified in natural language by exploring the environment. Current ObjectNav methods focus on prompt engineering for perception and do not address the semantic uncertainty induced by variations in prompt phrasing. Ignoring semantic uncertainty can lead to suboptimal exploration, which in turn limits performance. Hence, we propose a semantic uncertainty-informed active perception pipeline for ObjectNav in indoor environments. We introduce a novel probabilistic sensor model for quantifying semantic uncertainty in vision-language models and incorporate it into a probabilistic geometric-semantic map to enhance spatial understanding. Based on this map, we develop a frontier exploration planner with an uncertainty-informed multi-armed bandit objective to guide efficient object search. Experimental results demonstrate that our method achieves ObjectNav success rates comparable to those of state-of-the-art approaches, without requiring extensive prompt engineering.**

I. Introduction

Robots benefit from effective goal-directed exploration in human-centric environments to accomplish tasks such as locating and delivering objects, as well as assisting with household activities. These tasks require a robot to detect objects and understand spatial layouts with task-relevant contextual semantic cues. To this end, grounding language instructions in perceptual and spatial representations of the scene, also known as open-vocabulary perception, is key to efficient exploration and success of the task. However, open-vocabulary perception is inherently uncertain due to the wide variety of objects that exist in households and the many ways they can be described in natural language. Quantifying this perception uncertainty and accounting for it during exploration is an important element to realise reliable task execution.

This work examines the object goal navigation (ObjectNav) task [2], in which a robot must locate and navigate to a target object in an unfamiliar indoor environment within a limited mission time. The robot needs to recognise arbitrary and diverse objects based on a natural language description. The environment is unknown prior to a mission and is only partially observable due to noisy RGB and depth sensor observations, limited in range. To this end, the robot must continuously refine its exploration strategy online using its current sensor observations to navigate to the target object.

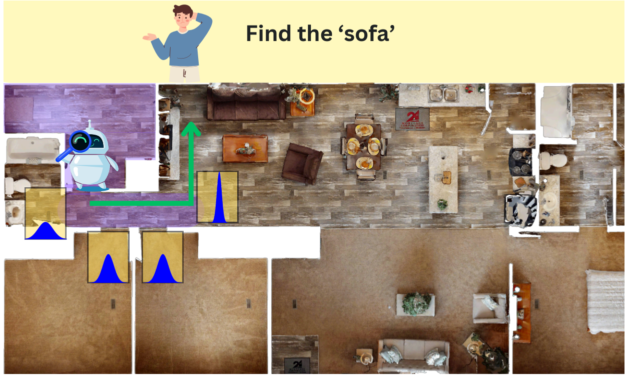

Fig. 1: We develop an uncertainty-informed, open-vocabulary ObjectNav pipeline for locating arbitrary objects in indoor environments. The figure visualises our approach: given a target object, the robot needs to navigate to it (green arrow) in an initially unknown environment. The robot actively selects a frontier to explore at each timestep amongst all available frontiers (yellow rectangles) using our multi-arm bandit frontier planner informed by semantic relevance estimates about each frontier (blue Gaussians) from our probabilistic geometric-semantic map (purple).

A common strategy employed by ObjectNav methods is to construct a 2D map of the environment online using observed geometric information, such as spatial layout and obstacles. Geometric maps enable tracking explored areas and guiding exploration toward map corners [15] or frontiers between known and unknown space, known as frontier-based exploration [27]. Semantic information, such as object categories, can also be integrated into maps using deep learning-based semantic segmentation [4], which enables exploration of areas with semantic similarity to the target object. However, most approaches are closed-vocabulary, i.e. limited to a fixed set of objects defined at training time and do not scale well to the wide variety of objects found in the real world. State-of-the-art approaches for ObjectNav [5], [7], [16], [29] leverage open-vocabulary perception using vision-language models (VLMs) [12], [18], allowing for semantically informed exploration scalable to potentially infinite object categories. Trained on large-scale image-text pairs, VLMs provide semantic relevance between arbitrary user-specified text prompts and images. However, VLMs are often sensitive to prompt phrasing. Variations in prompts,

Utkarsh Bajpai, Julius Rückin, Cyrill Stachniss are with the Center for Robotics, University of Bonn, Germany. Marija Popović is with the MAVLab, TU Delft, Netherlands. Cyrill Stachniss is also with the Lamarr Institute for Machine Learning and Artificial Intelligence, Germany.

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob and by the German Federal Ministry of Education and Research (BMBF) in the project "Robotics Institute Germany", grant No. 16ME0999. Corresponding author: utkarsh@uni-bonn.de.

such as "chair," "a chair," or "there is a chair nearby" produce different semantic relevance scores from VLMs, given the same image. This can lead to inconsistency in navigating to the target object in the ObjectNav scenario. Existing approaches [5], [29] often rely on fixed hand-tuned prompts appended to object names, which ignore this inconsistency and do not address this semantic uncertainty in VLM predictions. Ignoring semantic uncertainties is suboptimal and does not exploit the full potential of planning algorithms. Our paper aims to overcome this within the ObjectNav problem.

The main contribution of this paper is a semantic uncertainty-informed active perception framework for ObjectNav illustrated in Fig. 1. We propose a probabilistic geometric-semantic mapping method that updates geometric and semantic environment information online, while explicitly modelling the uncertainty in VLM-predicted semantic relevance. We propose a probabilistic sensor model using prompt ensembling [11], i.e. querying the VLM with varied prompts to approximate uncertainty in semantic relevance. Based on our probabilistic geometric-semantic map, we introduce a novel frontier planner with a multi-armed bandit objective for efficient exploration in ObjectNav. Our planner balances exploration of uncertain regions and exploitation of regions likely to contain the target object. By accounting for uncertainty in perception and planning, our framework eliminates the need for hand-tuned prompts.

In sum, we make the following claims. First, the success of prior open-vocabulary ObjectNav approaches relies on extensive prompt engineering, with the choice of prompts impacting downstream navigation performance. Second, we show that semantic relevance scores from a VLM prompt-ensemble are approximately normally distributed to validate the design of our probabilistic sensor model. Third, our uncertainty-informed planner performs comparably to state-of-the-art open-vocabulary ObjectNav approaches that rely on a single fixed, hand-tuned prompt. We provide open-source code at: https://github.com/PRBonn/uiap-ogn.

II. Related Work

Approaches for ObjectNav include end-to-end learning methods and modular methods. End-to-end learning methods [13], [26], [28] map observations directly to actions, training visual representation and navigation policies simultaneously in simulation environments. While successful in simulation, these methods encounter challenges in real-world deployment due to sample inefficiency and their reliance on a limited set of closed-world vocabulary object categories. This limits their ability to semantically relate and detect novel objects, as well as adapt to unseen environments.

In contrast, modular ObjectNav methods [5], [7], [29], [30], [32], [33] decompose the task into perception, mapping, exploration and point-goal navigation. Approaches build maps of the environment using geometric [22] and semantic information [6]. A planner selects waypoints for exploration of the environment, while point-goal navigation determines the actions required to reach them.

Recently, VLMs have significantly advanced robotic systems by enabling open-vocabulary object recognition and semantic reasoning. Early works such as VLMaps [8] leverage VLM-derived latent features to build 2D geometric-semantic maps by projecting them into a 3D geometric representation. However, the feature fusion process in VLMaps is computationally expensive, making it unsuitable for real-time applications. Other approaches, including scene graph-based methods [9], [24], construct hierarchical representations of semantic relationships and offer improved efficiency. Nevertheless, most scene graph representations remain too costly for real-time open-vocabulary exploration.

To enhance classical frontier-based exploration [27], recent ObjectNav methods incorporate open-vocabulary perception using VLMs. Clip on Wheels (CoW) [7] was among the first to use CLIP-derived features [18] for exploring frontiers until a target object is detected. Zhou et al. [33] use a CLIP-based object detector to detect objects and rooms and then use a language model to decide which room to navigate to. Recent works such as those by Chen et al. [5] and VLFM [29] incorporate VLM-derived cosine similarity scores between the recorded camera image and a fixed text prompt. VLFM projects these scores onto a 2D grid to generate a semantic relevance map. This approach eliminates the need for expensive feature fusion by updating a single scalar value per cell at each timestep.

Despite these advances, existing open-vocabulary ObjectNav approaches typically treat VLM-based semantic relevance predictions as point estimates, ignoring the semantic uncertainty inherent in natural language and visual observations. ObjectNav methods, such as VLFM [29], manually craft a fixed text prompt to elicit model responses, e.g. from BLIP-2 [12], which result in strong ObjectNav performance. These approaches ignore the underlying semantic uncertainty of VLMs, although minor variations in prompt phrasing cause differences in semantic relevance, resulting in inconsistent navigation behaviours and performance. Notably, prompt engineering is hard to scale and brittle across environments.

To address this, we propose an ObjectNav framework that explicitly models semantic uncertainty in open-vocabulary perception and integrates it into the planning process. Instead of relying on fixed prompts and treating similarity scores as point estimates, our method samples multiple prompt variants to approximate the distribution of similarity scores, thereby capturing semantic uncertainty using a new probabilistic sensor model. We fuse these probabilistic semantic relevance scores into an online-built geometric-semantic map. Our frontier-based planner leverages this map to guide exploration, dynamically balancing semantic relevance and its associated uncertainty. In this way, our ObjectNav framework improves robustness to linguistic ambiguities and enables informed planning in complex environments.

III. Problem Formulation

We formally define the ObjectNav task as an open-vocabulary, goal-directed navigation task [2]. In ObjectNav, a robot is deployed in an unknown environment $E \subseteq \mathbb{R}^3$.

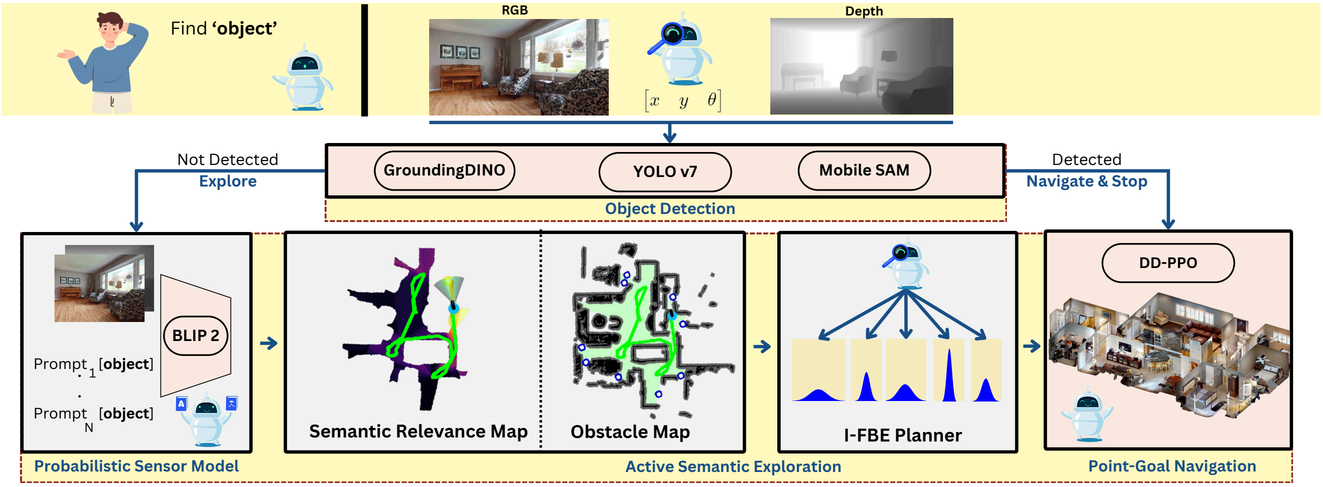

Fig. 2: Our ObjectNav approach consists of the object detection and active semantic exploration modules. If the target object is detected in the currently recorded frame, the robot navigates directly to it using point-goal navigation, as discussed in Sec. IV-A. Otherwise, it explores the environment using our uncertainty-informed frontier planner guided by our probabilistic semantic relevance map. Incorporating semantic cues and uncertainty into our map allows us to intelligently explore regions with higher probability of finding the target object.

The robot starts at an arbitrary pose $\mathbf{x}_0^w = (\mathbf{v}_0^w, r_0^w)^\top$, where $\mathbf{v}_0^w \in \mathbb{R}^3$ and $r_0^w \in SO(2)$ are the robot's starting position and rotation in the world coordinate frame w. The robot's goal is to navigate to an object in the environment, e.g. a cup, referred to as the "target object" $o \in \mathcal{O}$, where $\mathcal{O} \subset \mathbb{N}$ is a possibly infinite set of object categories present in the environment. The set of all target object instances in the environment is denoted as $\mathcal{G} = \{\mathbf{x}^w \in E \mid f(\mathbf{x}^w) = o\}$, where $f : E \rightarrow \mathcal{O}$ assigns an object category to each position in the environment. In each episode, the robot is tasked to navigate to a target object position $\mathbf{x}_g^w \in \mathcal{G}$. An episode $\tau = (E, o, \mathbf{x}_0^w)$ is defined by the a priori unknown environment E, user-defined target object category o, initial robot pose $\mathbf{x}_0^w$, and has a maximum length of $T \in \mathbb{N}$ timesteps. At each timestep t, the robot executes an action $a_t \in \mathcal{A}$, i.e. moving forward, turning left or right on the spot, or stopping. The robot receives an egocentric visual observation $\mathbf{I}_t = (\mathbf{I}_t^{\text{rgb}}, \mathbf{I}_t^{\text{depth}})$ from a RGB-D camera, where $\mathbf{I}_t^{\text{rgb}} \in \mathbb{R}^{U \times V \times 3}$ is the RGB image and $\mathbf{I}_t^{\text{depth}} \in \mathbb{R}^{U \times V}$ is the depth image with resolution $U \times V$. The robot pose $\mathbf{x}_t^w$ at each timestep is assumed to be known. If the robot's distance $\|\mathbf{x}_t^w - \mathbf{x}_g^w\|_2$ to the target object position $\mathbf{x}_g^w \in \mathcal{G}$ is smaller than a clearance distance $c > 0$, $\mathbf{x}_g^w$ is visible from $\mathbf{x}_t^w$, and the robot stops, then the episode is successful.

IV. Our Approach

Our uncertainty-informed ObjectNav approach explores the environment based on probabilistic estimates of semantic relevance. Our approach enables a robot to navigate to user-specified objects in an unknown 3D environment based on open-vocabulary perception, conceptually depicted in Fig. 2. We first perform object detection in each camera frame as described in Sec. IV-A. If the target object is detected, the robot directly navigates to it. Otherwise, we employ an active exploration strategy that identifies semantically relevant regions, such as a "kitchen" for target objects like "cup", by exploiting semantic correspondences and accounting for uncertainty using our probabilistic sensor model described in Sec. IV-B. We construct a geometric-semantic grid map based on the new sensor model, updating occupancy and semantic relevance information online, as discussed in Sec. IV-C. Our multi-arm bandit frontier-based planner uses our geometric-semantic map to guide the robot to regions with high semantic relevance, as outlined in Sec. IV-D.

A. Object Detection and Point Goal Navigation

We follow the approach of Yokohama et al. [29] for object detection. Each recorded RGB image is processed by the object detection module, which attempts to detect the target object $o \in \mathcal{O}$ in the frame using two object detectors. We rely on YOLOv7 [23], a supervised learning-based detector trained on a pre-defined closed set of 91 object categories $\mathcal{O}$, including a wide range of outdoor and indoor scenes. To allow our approach to work with potentially infinite and a priori unknown user-defined object categories of interest, we further use Grounding DINO [14] as an open-vocabulary object detector capable of detecting a wide range of objects specified by arbitrary text prompts. If the target object o is detected, we use MobileSAM [31] to segment o in the RGB image. To estimate the goal pose $\mathbf{x}_g^w$, the segmentation mask is applied to the depth image and projected into 3D world coordinates using known camera parameters. We use the centroid of the projected depth information of o to determine a target object position $\mathbf{x}_g^w \in \mathcal{G}$. Then, we use DD-PPO [25] to navigate to $\mathbf{x}_g^w$ from the current pose of the robot $\mathbf{x}_t^w$.

B. Probabilistic Semantic Relevance Sensor Model

To extract semantic uncertainty from VLMs, we introduce our probabilistic sensor model. In our active semantic exploration module, we use the BLIP2 [12] VLM for semantically guided exploration. VLMs project both visual and textual inputs into a shared latent embedding space.

Semantic relationships are preserved within this space, such that contextually similar images $\mathbf{I}^{\text{rgb}}$ and text prompts p are embedded as latent representations $\mathbf{l}^{\text{rgb}} \in \mathbb{R}^D$ and $\mathbf{l}^p \in \mathbb{R}^D$ respectively, oriented in similar directions. We leverage the commonly used cosine similarity metric to quantify the alignment between text and image embeddings, providing an estimate of semantic relevance between an image and text:

$$S(\mathbf{l}^{\text{rgb}}, \mathbf{l}^p) = \frac{\mathbf{l}^{\text{rgb}} \cdot \mathbf{l}^p}{\|\mathbf{l}^{\text{rgb}}\|_2 \|\mathbf{l}^p\|_2} \,. \tag{1}$$

However, VLMs are often sensitive to slight prompt variations, leading to fluctuating semantic relevance as measured in Eq. (1). We model the VLM's data uncertainty about semantic relevance as a Gaussian-distributed random variable

$$\mathcal{S} \;\sim\; \mathcal{N}(\mu_Z, \sigma_Z^2) \,, \tag{2}$$

where μ_Z is the mean semantic relevance score and σ_Z^2 its variance. We approximate the mean and variance of $\mathcal{S}$ in a Monte Carlo fashion by designing a prompt ensemble [11] using a set of N prompts $\mathcal{P} = \{p_1, p_2, \dots, p_N\}$. We generate these prompts using the GPT4 [1] language model, asking for alternative prompt formulations to the default prompts "there is a `target_object` ahead" and "A `target_object` is in the vicinity". The RGB image $\mathbf{I}^{\text{rgb}}$ and each prompt $p_i \in \mathcal{P}$ are mapped to their latent representations $\mathbf{l}^{\text{rgb}}$ and $\mathbf{l}^{p_i}$ by the VLM. Then, the mean μ_Z of the semantic relevance scores and their variance σ_Z^2 are:

$$\mu_Z = \frac{1}{|\mathcal{P}|} \sum_{p_i \in \mathcal{P}} S(\mathbf{l}^{\text{rgb}}, \mathbf{l}^{p_i}) \,, \tag{3}$$

$$\sigma_Z^2 = \frac{1}{|\mathcal{P}|} \sum_{p_i \in \mathcal{P}} \left(S(\mathbf{l}^{\text{rgb}}, \mathbf{l}^{p_i}) - \mu_Z \right)^2 . \tag{4}$$

Usually, not all parts of the sensor's field of view (FOV) contribute equally to semantic relevance. VLMs tend to assess semantic relevance more reliably for objects near the image centre, while peripheral regions often exhibit diminished influence on the semantic relevance. Following Yokohama et al. [29], we address this limitation by introducing a viewpoint-dependent confidence measure C_V. However, unlike their approach of using the confidence measure in the map update, we use it in our sensor model. For image pixels close to the optical axis, C_V is high, while it decreases towards the peripheral regions of the image:

$$C_V(\theta) = \cos^2\left(\frac{2\theta\pi}{\theta_{\text{fov}}}\right) \,, \tag{5}$$

where θ is the angle between the pixel and the optical axis, and θ_{fov} is the horizontal FOV of the robot's camera.

We combine this confidence measure with our image-based semantic relevance variance σ_Z^2 in Eq. (4) into a per-pixel semantic relevance variance $\sigma_Z^2(\theta)$:

$$\sigma_Z^2(\theta) = \sigma_Z^2 + (1 - C_V(\theta)) \,. \tag{6}$$

We use our new sensor model for semantic relevance in ObjectNav to update our probabilistic geometric-semantic map as detailed in the following Sec. IV-C.

C. Probabilistic Geometric-Semantic Mapping

Our geometric-semantic mapping stores and updates the belief about the initially unknown environment E based on new incoming sensor observations. We maintain two grid maps, updating the geometric and semantic environment information received through depth sensor observations and semantic relevance estimation as described in Sec. IV-B. The environment is discretised into a 2D top-down grid G of resolution $H \times W$. At each timestep t, the geometric obstacle map $\mathcal{M}_{O,t} : G \rightarrow \{0,1\}^{H \times W}$ is updated based on the depth image's $\mathbf{I}_t^{\text{depth}}$ focal cone projected onto the 2D plane using occupancy mapping [17], [29].

We update our novel semantic relevance map $\mathcal{M}_{S,t} : G \rightarrow [0,1]^{H \times W}$ using our semantic relevance sensor model in Eq. (2). We assume the prior belief $\mathcal{M}_{S,0}(m) \sim \mathcal{N}(\mu_{S,0}(m), \sigma_{S,0}(m))$ of a grid cell $m \in G$ to be normally distributed with an uninformed prior mean $\mu_{S,0} = 0.5$ and large variance $\sigma_{S,0}^2 = 0.5$, expressing that the environment is initially unknown. This choice of map prior enables us to formulate the posterior semantic relevance map update at timestep t in closed form as the prior is conjugate, given the normally distributed sensor model $\mathcal{S}$. At timestep t, we update the map $\mathcal{M}_{S,t}(m)$ for each grid cell in the camera's projected 2D focal cone:

$$\mu_{S,t}(m) = \frac{\sigma_{S,t-1}^2(m)\,\mu_{Z,t} + \sigma_{Z,t}^2(\theta)\,\mu_{S,t-1}(m)}{\sigma_{S,t-1}^2(m) + \sigma_{Z,t}^2(\theta)} \,, \tag{7}$$

$$\sigma_{S,t}^2(m) = \frac{\sigma_{S,t-1}^2(m)\,\sigma_{Z,t}^2(\theta)}{\sigma_{S,t-1}^2 + \sigma_{Z,t}^2(\theta)} \,, \tag{8}$$

where $\mathcal{S}_t \sim \mathcal{N}(\mu_{Z,t}, \sigma_{Z,t}^2(\theta))$ is the current semantic relevance measurement computed as in Eq. (3) and Eq. (6). As in the geometric obstacle map update, the semantic relevance measurement $\mathcal{S}_t$ is projected to the 2D plane using the camera's focal cone. Hence, we use $\mu_{Z,t}$ to update all grid cells m in the focal cone while $\sigma_{Z,t}^2(\theta)$ is the variance of a pixel described by θ that is projected onto grid cell m. In the following Sec. IV-D, we detail how we leverage the geometric-semantic map in our new ObjectNav planning method to explore the environment towards the target object of interest in a semantically-targeted fashion.

D. Uncertainty-Informed Exploration

Building on the probabilistic geometric-semantic map described in Sec. IV-C, we propose a new uncertainty-informed planner to perform semantically-targeted exploration towards the target object of interest. Our planner frames frontier exploration [27] as a multi-armed bandit problem. The planner seeks to find the next-best frontier to navigate towards by balancing the exploitation of frontiers with known high semantic relevance with exploration of semantically uncertain frontiers. We treat available frontiers as arms, evaluating their potential utility towards finding the target object by employing decision-theoretic reward functions.

At each timestep t, we leverage our geometric obstacle map $\mathcal{M}_{G,t}$ to navigate to frontiers of known free space and unexplored space. Let $m_i \in G$ be a grid cell representing the center of a frontier $i \in \{1, 2, \ldots, F_t\}$, where F_t is the number of frontiers in $\mathcal{M}_{G,t}$. We evaluate our current semantic relevance map belief $\mathcal{M}_{S,t}(m_i) \sim \mathcal{N}(\mu_{S,t}(m_i), \sigma^2_{S,t}(m_i))$ at each frontier i. To guide the robot's exploration, we develop two planners, called I-FBE1 and I-FBE2.

I-FBE1 uses the expected improvement function [10] to select a frontier to explore among all available frontiers based on their semantic relevance and uncertainty, as:

$$\begin{aligned} \mathrm{EI}(m_i) = \; & (\mu_{S,t}(m_i) - \mu_t^*)\, \Phi\left(\frac{\mu_{S,t}(m_i) - \mu_t^*}{\sigma_{S,t}(m_i)}\right) \\ & + \sigma_{S,t}(m_i)\, \phi\left(\frac{\mu_{S,t}(m_i) - \mu_t^*}{\sigma_{S,t}(m_i)}\right), \qquad (9) \\ \mu_t^* = \; & \max_{j \in \{1,\ldots,F_t\}} \mu_{S,t}(m_j)\,, \qquad (10) \end{aligned}$$

where μ_t^* is currently the highest semantic relevance among all frontiers in $\mathcal{M}_{G,t}$, and Φ and ϕ are the standard normal distribution's cumulative distribution and probability density function. The term $\Phi(\cdot)$ reflects the probability that the candidate frontier will improve upon the current best, supporting exploitation, while $\phi(\cdot)$ promotes exploration of uncertain frontiers that may lead to high potential improvement.

I-FBE2 uses the Gaussian process upper confidence bound (GP-UCB) [21] reward function. GP-UCB, in contrast with expected improvement, promotes aggressive exploration of frontiers, even if they are not promising for immediate gains:

$$\text{GP-UCB}(m_t^i) = \mu_{S,t}(m_t^i) + \sqrt{\beta}\sigma_{S,t}(m_t^i)\,. \qquad (11)$$

The first term $\mu_{S,t}(m_t^i)$ promotes exploitation by favoring regions with high predicted relevance, while the second term $\sqrt{\beta}\,\sigma_{S,t}(m_t^i)$ encourages exploration in areas of high uncertainty. The hyperparameter β controls the trade-off.

V. Experimental Results

We design our experiments to show the capabilities of our method. The results of our experiments support our key claims: (i) The success of prior open-vocabulary ObjectNav approaches relies on extensive prompt engineering, and prompt choice directly influences ObjectNav performance; (ii) We show that semantic relevance scores from a VLM prompt-ensemble are approximately normally distributed to validate the design of our probabilistic sensor model; (iii) Our uncertainty-informed planner performs comparably to the state-of-the-art open-vocabulary ObjectNav approaches that rely on fixed hand-engineered prompts.

A. *Experimental Setup*

Datasets and hardware. To assess the effectiveness of our ObjectNav approach, we conduct experiments in realistic indoor environments using the Habitat simulator [20], a popular simulator for ObjectNav evaluations [2]. We use two datasets, Matterport3D (MP3D) [3] and Habitat-Matterport 3D (HM3D) [19], both offering complex 3D reconstructions of indoor spaces. MP3D provides 3D scans with rich semantic annotations but limited unique environments and data imperfections. HM3D offers 1,000 high-quality scans with complex layouts and varied conditions. We evaluate our approach across varied, realistic scenes in 2,195 MP3D and 2,000 HM3D episodes, which contain 20 scenes, six object categories and 11 scenes and 21 object categories, respectively, following prior work [29]. Experiments are conducted on a workstation with a 12th-generation Intel Core i7 CPU, 64 GB RAM, and an NVIDIA RTX A5000 GPU.

Prompt	**SR ↑**	**SPL ↑**
Seems like there is a `target_object` ahead	52.60	30.42
A place where `target_object` can be found	51.00	29.71
A `target_object` can be in the vicinity	53.20	31.20
Seems like a `target_object` is ahead	53.20	30.50
A `target_object` is in the vicinity	51.65	28.67
`target_object` likely ahead	52.45	29.86
`target_object`	50.60	28.28
Ours (I-FBE1)	52.25	28.96
Ours (I-FBE2)	53.50	27.31

TABLE I: Impact of prompt phrasing variations on success rate (SR) and SPL in downstream navigation tasks. The prompt formulation significantly influences the performance on the ObjectNav task on the HM3D Dataset. Our uncertainty-informed methods perform comparably to the default prompt in VLFM [29] (top).

Evaluation metrics. We evaluate each approach using Success Rate (SR) and Success Weighted by Path Length (SPL) [2]. SPL measures the efficiency of the robot's path by comparing it to the shortest possible route from the starting point to the nearest instance of the target object in the ground truth. If the agent fails to reach the target, the SPL score is zero; otherwise, it represents the ratio of the shortest path length to the agent's actual path length, with higher values indicating better path efficiency.

Baselines. To benchmark our planners I-FBE1 and I-FBE2, we compare our pipeline with geometric variants of the frontier planner [27]: Closest-FBE and RandomFBE. Closest-FBE directs the robot to the nearest frontier for exploration, while Random-FBE selects frontiers at random. We also evaluate against open-vocabulary semantically informed approaches, VLFM [29], CoW [7], ESC [33], and ZSON [16]. ZSON projects the image of the target object and the prompt into the same embedding space using the CLIP [18] VLM and performs ObjectNav using a trained planning network. CoW uses gradient-based visualisation of the CLIP VLM along with frontier-based exploration, whereas ESC uses a VLM with a language model to guide frontier exploration. VLFM is an open-vocabulary frontier planner that uses VLM-derived cosine similarities to identify the most relevant frontiers. VLFM does not account for the uncertainty inherent in the VLM predictions and relies on a fixed prompt "Seems like there is a `target_object` ahead", making it a relevant baseline for evaluating the benefits of explicitly modeling semantic uncertainty.

B. *Effect of Prompt Phrasing on ObjectNav Performance*

Our first experiment is designed to demonstrate that the success of prior open-vocabulary ObjectNav approaches heavily depends on prompt engineering. In particular, we

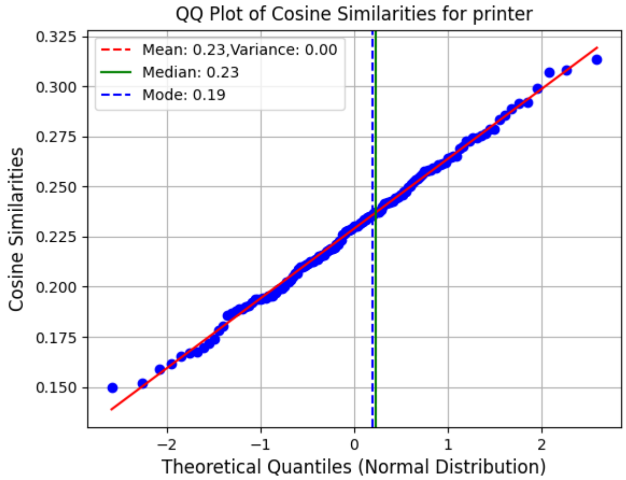

Fig. 3: We display the Quantile-Quantile (QQ) plot of 100 VLM-predicted semantic relevance scores, which are generated from 100 unique prompts around the target object name "printer". The quantiles of the semantic relevance scores (vertical) are contrasted with the theoretical quantiles of the standard normal distribution (horizontal). The linear relationship suggests that the semantic relevance scores are approximately normally distributed.

show that the choice of prompt directly impacts ObjectNav performance. We evaluate VLFM [29], which uses a fixed prompt: "Seems like there is a `target_object` ahead". We provide this prompt to the GPT4 [1] model and ask it to generate alternative formulations for the same target object, out of which we then use seven prompts to run ObjectNav evaluations. To assess performance, we only modify the prompt for the BLIP2 VLM in VLFM and evaluate on 2000 episodes of HM3D using the SR and SPL metrics.

Our results are presented in Tab. I. The choice of prompt for the VLM to estimate semantic relevance significantly impacts the overall effectiveness of the ObjectNav pipeline. For instance, a slight modification—changing the prompt to "Seems like a `target_object` is ahead" improved SR by 0.6%, while using only the object name reduced it by nearly 2% compared to the default hand-engineered prompt. These results highlight that minor prompt adjustments can lead to a difference in capturing semantic relevance, potentially adversely affecting ObjectNav performance. In contrast, our uncertainty-informed approaches I-FBE1 and I-FBE2 leverage all seven prompts to estimate semantic relevance. Particularly, I-FBE2 achieves performance close to the hand-crafted prompt in VLFM without manual prompt tuning.

C. Prompt Ensembling for Semantic Relevance

Our second experiment demonstrates that semantic relevance scores obtained from VLMs follow approximately a Gaussian distribution. This experiment validates the design of our probabilistic sensor model and the subsequent updates to its semantic relevance map. We use BLIP-2 [12] across five images and multiple object categories, including printer, oven, fridge, and table, with the target object category present in some images and absent in others. We generate prompt ensembles ranging from 10 to 100 variations for each category using GPT-4 [1]. We then compute the cosine similarity between each prompt and image as in Eq. (1).

Approach	HM3D		MP3D	
	SR↑	SPL↑	SR↑	SPL↑
Closest-FBE	11.80	9.34	-	-
Random-FBE	37.30	23.32	-	-
ZSON	25.50	12.60	15.30	4.80
CoW	-	-	7.40	3.70
ESC	39.20	22.30	28.70	14.20
VLFM	52.60	**30.40**	**36.40**	**17.50**
Ours (I-FBE1)	52.25	28.96	35.26	16.47
Ours (I-FBE2)	**53.50**	27.31	35.63	16.52

TABLE II: We compare our approach with the baselines Closest-FBE, Random-FBE, VLFM and other recent approaches for ObjectNav on the HM3D and MP3D datasets. Our I-FBE1 and I-FBE2 perform comparably to the state-of-the-art ObjectNav methods.

We visualise the distribution of semantic relevance scores for the printer category in Fig. 3 as a representative example. The results show that cosine similarities of prompt ensembles approximate a Gaussian distribution, as illustrated by the Quantile-Quantile (QQ) plot. These observations validate the design and assumptions of our sensor model in Sec. IV-B.

D. ObjectNav Performance

Our third experiment is designed to show that our uncertainty-informed planner performs comparably to state-of-the-art open-vocabulary ObjectNav approaches that rely on fixed hand-engineered prompts. We evaluate the two variants of our uncertainty-informed planner for the ObjectNav task across both MP3D and HM3D datasets and compare them to state-of-the-art approaches described in Sec. V-A.

Our results are summarized in Tab. II. Frontier-based Closest-FBE planning performs worst, often trapping the robot in narrow, nearby regions. Frontier-based Random-FBE avoids this by selecting random frontiers, enabling more thorough exploration of unknown space. However, neither approach is semantically informed but purely geometric, resulting in suboptimal performance. Among the semantically-informed methods, in line with prior works, VLFM outperforms CoW, ESC, and ZSON, due to the strength of its underlying VLM and updating the map with cosine similarity values for each received camera frame. Our uncertainty-informed planners achieve success rates comparable to VLFM, despite not relying on hand-crafted prompts. Particularly, I-FBE2 outperforms VLFM across many prompt variations on HM3D as shown in Tab. I. We demonstrate that open-vocabulary ObjectNav methods relying on a single hand-tuned prompt are brittle. In contrast, our uncertainty-informed planners leveraging our probabilistic sensor model offer a more reliable and robust alternative. However, our method consistently underperforms slightly on the SPL metric compared to VLFM. This is primarily because our planner is designed to actively explore regions with high semantic relevance uncertainty. As a result, the agent occasionally replans toward uncertain regions even if they are not highly semantically relevant, leading to detours en route to the target object. On our evaluation setup, each replanning step takes approximately 420 ms of wall-clock time per timestep.

VI. Conclusions and Future Work

In this work, we proposed a training-free, open-vocabulary, semantic uncertainty-informed active perception pipeline for the ObjectNav problem in mobile robotics. We introduce a novel probabilistic sensor model based on prompt ensembles to estimate semantic relevance and uncertainty from VLMs. Our framework incorporates a probabilistic geometric-semantic map and uncertainty-informed frontier planners to address the brittleness of existing ObjectNav approaches, which rely on a single, fixed, hand-engineered prompt. Our experimental results show that the success of prior open-vocabulary ObjectNav relies on extensive prompt engineering. In contrast, our uncertainty-informed planner performs comparably to state-of-the-art open-vocabulary ObjectNav approaches that rely on fixed hand-engineered prompts while reducing dependence on prompt engineering. Future work will investigate the real-world deployment of our approach in larger environments under sensor noise.

References

[1] J. Achiam, S. Adler, et al. GPT-4 Technical Report. *arxiv preprint, arXiv:2303.08774*, 2024.

[2] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans. ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. *arXiv preprint, arXiv:2006.13171*, 2020.

[3] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *Proc. of Intl. Conf. on 3D Vision (3DV)*, 2017.

[4] D.S. Chaplot, D. Gandhi, A. Gupta, and R. Salakhutdinov. Object Goal Navigation using Goal-Oriented Semantic Exploration. In *Proc. of the Conf. on Neural Information Processing Systems*, 2020.

[5] J. Chen, G. Li, S. Kumar, B. Ghanem, and F. Yu. How To Not Train Your Dragon: Training-free Embodied Object Goal Navigation with Semantic Frontiers. In *Proc. of Robotics: Science and Systems*, 2023.

[6] F. Endres, C. Plagemann, C. Stachniss, and W. Burgard. Scene Analysis using Latent Dirichlet Allocation. In *Proc. of Robotics: Science and Systems*, 2009.

[7] S.Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song. CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023.

[8] C. Huang, O. Mees, A. Zeng, and W. Burgard. Visual Language Maps for Robot Navigation. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2023.

[9] N. Hughes, Y. Chang, and L. Carlone. Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. In *Proc. of Robotics: Science and Systems*, 2022.

[10] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[11] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[12] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proc. of the Intl. Conf. on Machine Learning*, 2023.

[13] Y. Li, A. Debnath, G.J. Stein, and J. Kosecka. Learning-Augmented Model-Based Planning for Visual Exploration. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2023.

[14] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[15] H. Luo, A. Yue, Z.W. Hong, and P. Agrawal. Stubborn: A Strong Baseline for Indoor Object Navigation. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2022.

[16] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. In *Proc. of the Conf. on Neural Information Processing Systems*, 2022.

[17] H. Moravec and A. Elfes. High resolution maps from wide angle sonar. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 1985.

[18] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proc. of the Intl. Conf. on Machine Learning*, 2021.

[19] S.K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J.M. Turner, E. Undersander, W. Galuba, A. Westbury, A.X. Chang, M. Savva, Y. Zhao, and D. Batra. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. In *Proc. of the Conf. on Neural Information Processing Systems*, 2021.

[20] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra. Habitat: A Platform for Embodied AI Research. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*, 2019.

[21] N. Srinivas, A. Krause, S.M. Kakade, and M.W. Seeger. Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Trans. on Information Theory*, 58(5):3250–3265, 2012.

[22] C. Stachniss, J. Leonard, and S. Thrun. *Springer Handbook of Robotics, 2nd edition*, chapter Chapt. 46: Simultaneous Localization and Mapping. Springer Verlag, 2016.

[23] C.Y. Wang, A. Bochkovskiy, and H.Y.M. Liao. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023.

[24] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard. Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation. *Proc. of Robotics: Science and Systems*, 2024.

[25] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra. DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames. In *Proc. of the Intl. Conf. on Learning Representations*, 2020.

[26] K. Yadav, R. Ramrakhya, A. Majumdar, V.P. Berges, S. Kuhar, D. Batra, A. Baevski, and O. Maksymets. Offline visual representation learning for embodied navigation. In *Workshop on Reincarnating Reinforcement Learning at ICLR*, 2023.

[27] B. Yamauchi. A Frontier-Based Approach for Autonomous Exploration. In *Proc. of the IEEE Intl. Symp. on Computer Intelligence in Robotics and Automation*, 1997.

[28] J. Ye, D. Batra, A. Das, and E. Wijmans. Auxiliary Tasks and Exploration Enable ObjectGoal Navigation. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*, 2021.

[29] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher. VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2024.

[30] B. Yu, H. Kasaei, and M. Cao. L3MVN: Leveraging Large Language Models for Visual Target Navigation. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2023.

[31] C. Zhang, D. Han, Y. Qiao, J.U. Kim, S.H. Bae, S. Lee, and C.S. Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arxiv preprint, arXiv:2306.14289*, 2023.

[32] L. Zhang, Q. Zhang, H. Wang, E. Xiao, Z. Jiang, H. Chen, and R. Xu. TriHelper: Zero-Shot Object Navigation with Dynamic Assistance. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2024.

[33] K. Zhou, K. Zheng, C. Pryor, Y. Shen, H. Jin, L. Getoor, and X.E. Wang. ESC: exploration with soft commonsense constraints for zero-shot object navigation. In *Proc. of the Intl. Conf. on Machine Learning*, 2023.