OpenFMNav: Towards Open-Set Zero-Shot Object Navigation via Vision-Language Foundation Models

Anonymous ACL submission

Abstract

Object navigation (ObjectNav) requires an agent to navigate through unseen environments to find queried objects. Many previous methods attempted to solve this task by relying on supervised or reinforcement learning, where they are trained on limited household datasets with close-set objects. However, two key challenges are unsolved: understanding free-form natural language instructions that demand open-set objects, and generalizing to new environments in a zero-shot manner. Aiming to solve the two challenges, in this paper, we propose **OpenFMNav**, an **Open**-set **F**oundation **M**odel based framework for zero-shot object **Nav**igation. We first unleash the reasoning abilities of large language models (LLMs) to extract proposed objects from natural language instructions that meet the user's demand. We then leverage the generalizability of large vision language models (VLMs) to actively discover and detect candidate objects from the scene, building a *Versatile Semantic Score Map (VSSM)*. Then, by conducting common sense reasoning on *VSSM*, our method can perform effective language-guided exploration and exploitation of the scene and finally reach the goal. By leveraging the reasoning and generalizing abilities of foundation models, our method can understand free-form human instructions and perform effective open-set zero-shot navigation in diverse environments. Extensive experiments on the HM3D ObjectNav benchmark show that our method surpasses all the strong baselines on all metrics, proving our method's effectiveness. Furthermore, we perform real robot demonstrations to validate our method's open-set-ness and generalizability to real-world environments.

1 Introduction

As a fundamental task in robotics and embodied AI, object navigation requires an agent to navigate through unseen environments to find queried objects. Compared to other robotic tasks, it is particularly important because it is a prerequisite for robots to interact with objects. To address this issue, several household datasets and benchmarks, such as MP3D (Chang et al., 2017), Gibson (Xia et al., 2018) and HM3D (Ramakrishnan et al., 2021) are proposed. Many previous studies (Chaplot et al., 2020a; Ramrakhya et al., 2022; Zhang et al., 2023) have attempted to solve this problem through supervised or reinforcement learning, where they are trained on particular household datasets above with close-set objects and comparable environments.

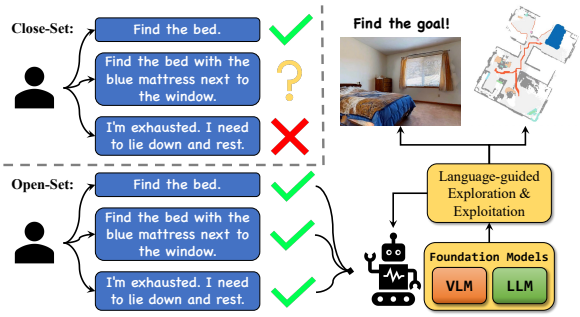


Figure 1: Leveraging foundation models, our proposed OpenFMNav can follow free-form natural language instructions with open-set objects and achieve effective zero-shot object navigation.

However, there are two significant challenges remaining unsolved. **First**, as shown in Fig 1, in many scenarios, instead of only mentioning an object category (e.g., *"Find the bed."*), humans often provide free-form instructions, either specifying objects with specific characteristics (e.g., *"Find the bed with the blue mattress next to the window."*), or expressing their demand without explicitly mentioning the object (e.g., *"I'm exhausted. I need to lie down and rest."*). These natural language instructions may demand open-set objects not included in the training vocabulary. In such cases, existing supervised or reinforcement learning-based methods fail to understand these natural language

instructions since they require specific object categories and were trained to perform close-set object detection. **Second**, due to the data scarcity of embodied navigation (Gu et al., 2022), these methods are typically trained on limited datasets that only cover household environments, which causes severe overfitting issues and prevents them from generalizing to unseen and diverse environments, let alone performing zero-shot navigation.

To address the first challenge, some initial progress has been made in understanding free-form natural language instructions with open-set objects. For instance, demand-driven navigation (DDN) was proposed by Wang et al. (2023) to map human instructions to a demand-conditioned attribute space. However, it is still limited to household settings and cannot be generalized to various environments. Another approach was suggested by Majumdar et al. (2023), which involves finding objects with specific attributes and eliminating distractors. However, it needs 2D occupancy maps and pre-exploration of the scene in the beginning, which are unavailable in unseen environments.

On the second challenge, recent years have witnessed progress in Zero-Shot Object Navigation (ZSON) (Majumdar et al., 2022; Gadre et al., 2023; Yokoyama et al., 2023; Zhou et al., 2023; Dorbala et al., 2023; Yu et al., 2023; Shah et al., 2023; Cai et al., 2023). However, some of these works (Majumdar et al., 2022; Yu et al., 2023; Cai et al., 2023) require data to train specific modules such as locomotion planning, and hence are not real "Zero-Shot". More importantly, these methods cannot conduct explicit and comprehensive reasoning on free-form natural language instructions, leading to their low performance and preventing them from being applied to many downstream robotic tasks.

To better address the aforementioned two key challenges, in this paper, we propose OpenFMNav, a novel framework based on foundation models to achieve effective open-set zero-shot navigation. To this end, we utilize foundation models to leverage their reasoning abilities and generalizability to interpret human instructions and actively explore the environment. To be more specific, we first leverage large language models to extract initially proposed objects from natural language instructions and merge them with user-defined prior objects and objects discovered by vision language models. We then construct an object prompt to detect and segment objects from the observation image, leveraging large vision language models. By using depth images to project the segmentation masks to the space, we can build a 2D top-down *Versatile Semantic Score Map (VSSM)* of the whole scene, on which we sample frontiers with semantic information for a large language model to conduct common sense reasoning and wisely choose frontiers to guide navigation. This way, we can perform language-guided exploration and exploitation of the scene and achieve effective open-set zero-shot object navigation without prior training on any household datasets. Moreover, unlike previous map-based methods such as Zhou et al. (2023); Yu et al. (2023); Shah et al. (2023); Yokoyama et al. (2023), the *VSSM* produced by our method will keep updating during the navigation, which better adapts to changing environments and can be further used in downstream robotic tasks, such as multi-goal navigation and mobile manipulation.

We conduct extensive experiments on the HM3D ObjectNav benchmark (Yadav et al., 2022a). Results show that our method outperforms the State-of-the-Art open-set zero-shot object navigation method (Zhou et al., 2023) by over 15% on success rate and surpasses all the strong baselines on all metrics, validating the effectiveness and superiority of our framework. Additionally, our method has been proven to understand free-form natural language instructions with open-set objects and generalize well to real-world environments through real robot demonstrations.

2 Related Work

2.1 Embodied Navigation

Embodied navigation is a fundamental yet challenging task in robotics and embodied AI since it is the precursor to many downstream robotic tasks, such as object manipulation and teleoperation. In such scenarios, given a specific goal and egocentric observations, agents are required to move to a desired location within a maximum timestep.

Due to the importance of embodied navigation, recent years have witnessed several branches of navigation tasks with different goal specifications. For instance, point goal navigation (PointNav) (Wijmans et al., 2019; Savva et al., 2019) uses point coordinates in the space as the goal; image goal navigation (ImageNav) (Chaplot et al., 2020b; Savinov et al., 2018) requires the agent to move where the given image is taken; and vision-language navigation (VLN) (Anderson et al., 2018; Ku et al., 2020) requires the agent to follow step-by-step in-

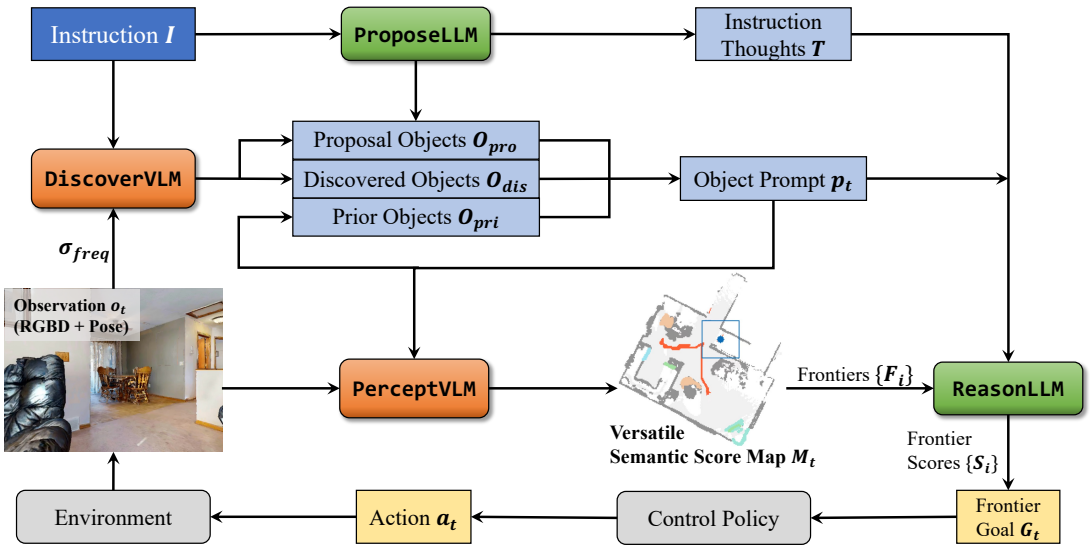


Figure 2: The framework of our proposed OpenFMNav. Based on the natural language instruction and observations, we utilize foundation models to interpret human instructions and construct a *Versatile Semantic Score Map (VSSM)*, on which we perform common sense reasoning and scoring to conduct language-guided frontier-based exploration.

structions to reach the location; and in object navigation (ObjectNav) (Batra et al., 2020), the agent is required to find objects of specified categories.

Compared to vision-language navigation (VLN), which offers detailed and step-by-step instructions, object navigation is particularly challenging since the agent needs to do semantic recognition to find the goal and needs more efficient exploration than VLN since there are no step-by-step instructions (Chen et al., 2023). It is also more common in real life that humans will give ambiguous demands (Wang et al., 2023) rather than detailed instructions in VLN. Additionally, VLN datasets (Anderson et al., 2018; Ku et al., 2020) are typically discretized into checker-like waypoint graphs, which makes it difficult to deploy algorithms in the real world, while on the other hand, ObjectNav is **continuous** and more **object-centric** so that it can be easily deployed and extended to many downstream robotic tasks like object manipulation.

To take a step further, in this paper, we propose a solution to the problem of open-set-ness in ObjectNav by introducing a framework that transforms the paradigm of ObjectNav from given close-set category names to free-form natural language instructions with open-set objects. This transformation will help bridge the interaction between humans and embodied agents, making it more useful in real-world applications. Compared to existing works (Majumdar et al., 2023; Wang et al., 2023), our method doesn't need prior occupancy maps and pre-exploration in the beginning and thus can navigate in unseen environments. Furthermore, our method addresses the overfitting issue in embodied navigation and easily generalizes to the real world in a zero-shot manner, enabling intelligent robot agents to navigate in more diverse environments.

2.2 Zero-Shot Object Navigation

As Gu et al. (2022) elaborates, embodied navigation faces a severe challenge of data scarcity, limiting the amount and distribution of available data for training. Methods directly supervised on these limited data cannot generalize to diverse real-world environments.

Therefore, recent years have witnessed great progress in Zero-Shot Object Navigation (ZSON). Methods proposed by Majumdar et al. (2022); Gadre et al. (2023); Yokoyama et al. (2023) leverage CLIP (Radford et al., 2021) or BLIP-2 (Li et al., 2023) embedded features to compute similarities between object goal and input image and construct an implicit map for certain goal objects to guide navigation. Other methods, such as those proposed by Zhou et al. (2023); Dorbala et al. (2023); Yu et al. (2023); Shah et al. (2023), leverage object detectors to construct metric maps and use large language models to conduct reasoning. Cai et al. (2023) leverages foundation models to perform basic image processing and trains a locomotion module to navigate to certain chosen pixel points.

2.3 Foundation Models

Foundation models (Bommasani et al., 2022) are large-scale models that are pre-trained on vast amounts of data and can perform general tasks. The sheer volume of pretraining data endows them with exceptional generalizability, which allows them to perform zero-shot inference. Moreover, the extensive training data helps foundation models acquire common sense about our physical world, making them ideal for real-world applications.

Foundation models, particularly the large language models (LLMs), also have an intriguing feature — In-Context Learning (ICL) (Dong et al., 2023). This feature enables these models to follow pre-defined instructions to ground their output into certain patterns. By combining ICL with common sense learned from the large-scale data, foundation models can effectively perform semantic reasoning and guesswork to provide intuitions of possible exploration directions like human beings, as illustrated in Zhou et al. (2023); Yu et al. (2023); Shah et al. (2023). For example, if the goal is a "toilet", from common sense it is highly possible to find it around an area that contains a "bathtub".

According to different modalities, foundation models can be mainly divided into Visual Foundation Models (VFM), such as SAM (Kirillov et al., 2023), Large Language Models (LLM), such as GPT-3.5/GPT-4 (Ouyang et al., 2022; OpenAI, 2023) and LLaMA/LLaMA-2 (Touvron et al., 2023a,b), and Vision Language Models (VLM), such as GPT-4V (Yang et al., 2023b), CLIP (Radford et al., 2021), Grounded-SAM (Liu et al., 2023), etc. There are also foundation models covering other modalities, such as audio (Yang et al., 2023a) and video (Xu et al., 2021). In this paper, we use VLMs and LLMs since our setting only involves vision and language modalities.

3 Method

3.1 Problem Statement and Method Overview

Problem Statement. As shown in Fig. 1, in an unfamiliar environment, given a natural language instruction I, an embodied agent needs to explore the environment in search of a certain queried object. At timestep t, the agent is provided with egocentric RGBD observation o_t and should output an action a_t such as `move_forward`, `turn_left`, `stop`, etc. A successful navigation is defined as finding the queried object within the maximum navigation timestep.

Method Overview. As shown in Fig. 2, given a starting point and human instruction I, the agent first utilizes the ProposeLLM to propose possible objects to meet the instruction. At timestep t, the agent can leverage the DiscoverVLM to discover new objects from the scene and check whether they can meet the instruction. Along with prior defined objects and proposal objects, the full object list is then converted into an object prompt p_t for foundation models to reason. Given current RGBD observation o_t, the PerceptVLM will detect and segment object masks based on p_t, constructing a *Versatile Semantic Score Map (VSSM)* M_t, on which possible exploration frontiers are sampled. Finally, the ReasonLLM will conduct common sense reasoning based on the semantic information of frontiers and give the next frontier goal G_t to explore, which will be executed by an underlying control policy to output low-level actions. The whole process is looped until the object is found or the agent fails.

3.2 Discovery and Perception

Discovery. Given a free-form human instruction I that may contain open-set objects, we first leverage a ProposeLLM to get all possible proposal objects O_{pro} that can satisfy the instruction. Each proposal object contains attributes such as color, location, etc, to satisfy fine-grained instructions. At timestep t, given egocentric RGBD and pose observations o_t, we propose a DiscoverVLM using GPT-4V (Yang et al., 2023b) that actively discovers novel objects O_{dis} from the RGB image. Meanwhile, the DiscoverVLM also conducts reasoning on the instruction, trying to discover objects that potentially meet the instruction and update O_{pro}. Extracting novel objects from the environment is essential for open-set navigation since they may contain scene-specific information that helps to find the goal. To save time and cost, the DiscoverVLM is randomly activated by a frequency parameter σ_{freq}.

Perception. After getting proposal objects O_{pro} and discovered objects O_{dis}, we merge them with prior objects O_{pri} to construct an object prompt p_t to feed into our PerceptVLM based on Grounded-SAM (Liu et al., 2023) to detect and segment all the appearing objects in p_t from the RGB image of o_t. Note that due to the BERT encoder (Devlin et al., 2019) and powerful SAM backbone (Kirillov et al., 2023) in the PerceptVLM, it can achieve open-set object detection in high granularities. This process will output object masks with confidence scores for further mapping and reasoning.

3.3 Mapping and Reasoning

Mapping. At timestep t, based on the confidence scores of object masks produced by PerceptVLM and the depth image and pose in o_t, we project the masks to the top-down 2D space and construct a *Versatile Semantic Score Map (VSSM)* $M_t \in \mathbb{R}^{H \times W \times (C+2)}$, which contains C channels of object semantics, and two channels of the occupied area and explored area, with a resolution of $H \times W$. Each element in the map is a score in $[0, 1]$ instead of binary labels. Since we continuously discover novel objects from the environment, the C is versatile so that we can keep updating the map, enabling life-long learning and downstream robotic tasks. Also, instead of filling binary labels into semantic channels, we fill each semantic channel with confidence scores, with which we can easily update the map if there is a change in the environment.

Reasoning. Based on M_t, we can sample frontiers $\{F_i\}$ with semantic information in unexplored areas for further exploration. To choose the next frontier to explore, we leverage ReasonLLM by unleashing the power of LLM's common sense reasoning. Specifically, given the semantic information around each frontier, we construct a query template in the form of "`This area contains A, B and C.`". Combined with the thought T produced by Chain-of-Thought (Wei et al., 2022) prompting from ProposeLLM and the object prompt p_t, the ReasonLLM will conduct common sense reasoning and rate these frontiers to pick one frontier goal G_t which is most likely to find the goal. This frontier goal G_t will guide the agent for further exploration and produce low-level actions to control the agent.

Instead of just asking the LLM which frontier to go to, we introduce another reasoning process, which prompts LLMs to rate these frontiers $\{F_i\}$ to scores $\{S_i\}$, in which $S_i \in [0, 1]$, indicating the likelihood to find the goal. Then, the frontier with the highest score will be picked out for further exploration. By introducing this rating process, ReasonLLM can map its common sense to concrete numbers that reflect the actual ranking, leading to better reasoning. We verified its effectiveness in Section 4.5. It's also worth mentioning that to balance exploration and exploitation, ReasonLLM is activated at regular timestep intervals δ to update G_t. At other timesteps, the frontier goal G_t remains unchanged to fully explore the previously chosen frontier $G_{t-\delta}$.

Algorithm 1: Pseudo-Code of the Overall Algorithm for OpenFMNav

```
Data: Natural Language Instruction I, Prior Objects
      O_pri, Discovery Frequency σ_freq, Frontier
      Goal Update Interval δ
t ← 0;
done ← False;
G_0, M_0, O_dis ← None;
O_pro, T ← ProposeLLM(I);
while not done do
    o_t ← getObservation();
    if toDiscover(σ_freq) then
        O_dis, O_pro ← DiscoverVLM(o_t, I);
    end
    p_t ← getPrompt(O_pro, O_dis, O_pri);
    Masks ← PerceptVLM(o_t, p_t);
    M_t ← semanticMapping(M_{t-1}, Masks, o_t);
    if O_pro in M_t then
        G_t ← getLocation(M_t, O_pro);
    else
        if t % δ == 0 then
            {F_i} ← sampleFrontiers(M_t);
            {S_i} ← ReasonLLM({F_i}, p_t, T);
            G_t ← getLocation(M_t, argmax({S_i}));
        else
            G_t ← G_{t-1};
        end
    end
    O_pri ← updateObj(O_pro, O_dis, O_pri);
    a_t ← FMMPlanner(M_t, G_t);
    done ← stepAction(a_t, t);
    t ← t + 1;
end
```

After obtaining the frontier goal and the occupancy channel in M_t, we utilize a control policy based on the Fast Marching Method (FMM) (Sethian, 1999) to output a low-level action a_t to control the agent. This closes the loop and goes to the next timestep $t + 1$.

We present the whole process of our OpenFMNav algorithm in Algorithm 1.

4 Experiments

In this section, we evaluate our method comprehensively in simulation to show our method's effectiveness compared to baseline methods. We also conducted ablation studies to validate the effectiveness of our framework design.

4.1 Experimental Setup

In the simulation, we evaluate on the HM3D ObjectNav benchmark based on the Habitat Matterport 3D Semantics Dataset (Yadav et al., 2022b), which contains 80 train scenes and 20 validation scenes. We utilize the validation scenes for evaluation. There are, in total, 2000 episodes and six goal

Method	Open-Set	Zero-Shot	SR (%) ↑	SPL ↑
FBE (Gervet et al., 2023)	×	✓	23.7	0.123
SemExp (Chaplot et al., 2020a)	×	×	37.9	0.188
ZSON (Majumdar et al., 2022)	✓	×	25.5	0.126
GoW (Gadre et al., 2023)	✓	✓	32.0	0.181
ESC (Zhou et al., 2023)	✓	✓	38.5	0.220
L3MVN (Yu et al., 2023)	×	✓	50.4	0.231
PixNav (Cai et al., 2023)	✓	×	37.9	0.205
OpenFMNav (Ours)	✓	✓	**54.9**	**0.244**

Table 1: Comparison between different methods on the HM3D ObjectNav benchmark. Our method outperforms all the baseline methods on all metrics and achieves open-set zero-shot object navigation.

classes (chair, couch, potted plant, bed, toilet, and tv) in the dataset. The action space of the robot agent is {`stop`, `move_forward`, `turn_left`, `turn_right`, `look_up`, `look_down`}. The forward distance is set to 0.25m, and the rotation angle is set to 30 degrees.

Following previous works (Zhou et al., 2023; Cai et al., 2023), we utilize **Success Rate (SR)** metric to measure whether an agent can find our desired objects. We also report results of **Success weighted by Path Length (SPL)** to measure the navigation efficiency.

4.2 Implementation Details

In our method, the foundation models we use are: GPT-4 (text-only) (OpenAI, 2023) for ProposeLLM and ReasonLLM, and GPT-4V (Yang et al., 2023b) for DiscoverVLM. For PerceptVLM, we utilize Grounded-SAM, which first leverages Grounding DINO (Liu et al., 2023) to produce bounding boxes given the RGB image in o_t and object prompt p_t, and then leverages Segment Anything Model (SAM) (Kirillov et al., 2023) for each bounding box to produce high-granularity object masks for semantic mapping.

Moreover, we utilize the Chain-of-Thought (CoT) (Wei et al., 2022) prompting technique to fully exploit the reasoning abilities of ProposeLLM, ReasonLLM and DiscoverVLM. The prompts we used can be found in Appendix C.

In the simulation, we set the update interval δ to 20 timesteps, discovery frequency σ_{freq} to 0.01, and the initial prior objects to a subset of HM3D object categories, which can be found in Appendix B.

4.3 Baseline Methods

We compare our method with several recent works, with a focus on open-set and zero-shot object navigation baselines to verify our framework's effectiveness. We classify these baseline methods into "Open-Set" and "Zero-Shot" or not. Here, we define "Open-Set" as that the method can find whatever object category we want, and define "Zero-Shot" as that the agent hasn't been trained or finetuned on ***any*** of the data previously, including images, episodes, and locomotion planning. The baseline methods are as follows:

- **FBE** (Gervet et al., 2023). This baseline method employs a classical robotics pipeline for mapping and a frontier-based exploration algorithm.
- **SemExp** (Chaplot et al., 2020a). A method that explores and searches for the target using close-set semantic maps and reinforcement learning.
- **ZSON** (Majumdar et al., 2022). An RGB-based zero-shot object navigation baseline using CLIP (Radford et al., 2021) to embed scene features. It is trained on ImageNav and directly transferred to ObjectNav.
- **GoW** (Gadre et al., 2023). A modification of CoW (Gadre et al., 2023) implemented by Zhou et al. (2023) that uses GLIP (Li* et al., 2022) for object detection and the vanilla fronter-based exploration method.
- **ESC** (Zhou et al., 2023). A map-based zero-shot object navigation baseline that uses GLIP (Li* et al., 2022) to detect objects and rooms, and combines LLM with soft commonsense constraints for planning.
- **L3MVN** (Yu et al., 2023). An LLM-based baseline that uses a close-set object detector (Jiang et al., 2018) and LLMs to conduct frontier-based exploration.

Method	SR (%) ↑	SPL ↑
w/o GPT-4	53.6	0.230
w/o CoT	51.8	0.208
w/o Discovery	50.0	0.222
w/o Scoring	50.0	0.208
Ours	**55.4**	**0.239**

Table 2: Ablation studies on different components of our method. Experiments are conducted under the same uniformly sampled episodes.

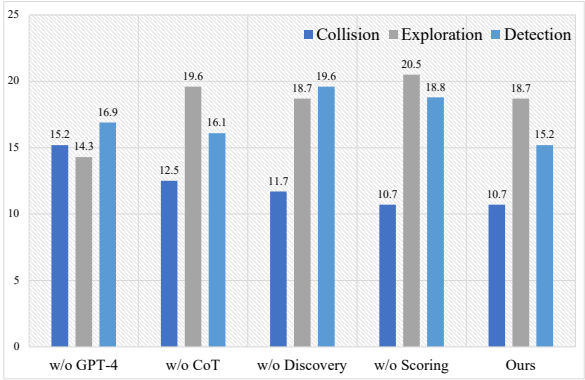


Figure 3: Types and percentages of failure cases in ablation methods.

- **PixNav** (Cai et al., 2023). A recent work solely using foundation models to pick out navigation pixels and train a locomotion module to navigate to chosen pixels.

4.4 Results and Analysis

We report the main results in Table 1. Our method surpasses all the baselines on both Success Rate (SR) and Success weighted by Path Length (SPL), especially compared with open-set zero-shot methods. Our method surpasses the previous State-of-the-Art method on open-set zero-shot object navigation (Zhou et al., 2023) by over 15% on the success rate metric, suggesting that our framework is indeed effective.

First, we compare our method with previous semantic map based methods, such as SemExp (Chaplot et al., 2020a), ESC (Zhou et al., 2023) and L3MVN (Yu et al., 2023). The results show that our method performs better since we utilize DiscoverVLM to construct *VSSM* with versatile out-of-vocabulary class labels, such as "marble statue" and "range hood", which helps to alleviate the issue of limited categories and enriches the semantic information of the environment. Also, compared to these methods, our method achieves open-set navigation, which better adapts to complex situations and real-world applications.

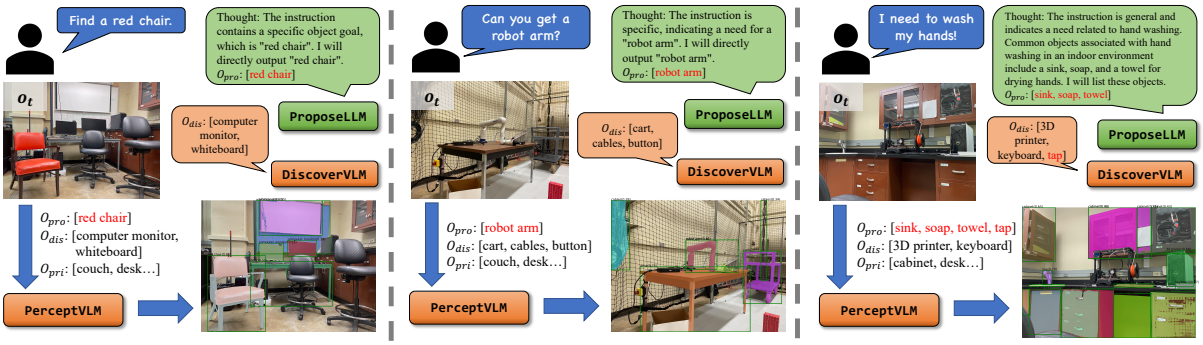
Compared with other open-set baselines, such as PixNav (Cai et al., 2023), ZSON (Majumdar et al., 2022) and GoW (Gadre et al., 2023), our method constructs an explicit map where all discovered objects are presented. Therefore, we can boost LLMs' reasoning abilities to balance between exploration and exploitation and make the agent move to where the goal is most likely to be. Also, the map constructed by our method is maintained and updated, which is perfect for life-long learning, enabling downstream robotic tasks with further natural language instructions, while methods like Gadre et al. (2023); Yokoyama et al. (2023) only construct implicit maps for a certain goal, which is useless after the navigation.

4.5 Ablation Studies

Probing deeper into our method design, we also performed ablation studies on various components of our pipeline. Note that to save time and cost, we test all the ablation methods on a subset of the full dataset under the same uniformly sampled episodes so that there can be slight differences in the result of our method. Table 2 shows that modifying multiple components of our framework leads to significantly worse performance. We also categorized the failure cases into different types and report their percentages in Fig. 3, in which **Collision** refers to the situation where the agent cannot avoid colliding with the environment, **Exploration** means the agent times out while trying to find the goal, and **Detection** means the agent mistakenly identifies a wrong object as the goal.

Effectiveness of using larger models. First, we analyze the usage of GPT-4 for LLMs. Compared to only using GPT-3.5, using larger GPT-4 achieves better performance (+1.8%), reducing failure cases of Collision and Detection. However, the percentage of Exploration is slightly higher, showing that larger models have more diverse answers that encourage more exploration, which potentially causes more time out.

Effectiveness of our joint reasoning pipeline. Then, we analyze different foundation model components. We found that using CoT prompting (+3.6%) and scoring prompting (+5.4%) are essential to the strong performance of OpenFMNav since they generate more reasoning chains that elicit the common sense of large language mod-



(a) Robust to distractors (b) Robust to open-set objects (c) Robust to free-form demands

Figure 4: Qualitative studies in the real world. Text marked in red indicates objects that potentially satisfy the instruction. Results show that our method is robust to natural language instructions, including distractors, open-set objects and free-form demands.

els. Also, compared to restricting the object set, leveraging DiscoverVLM not only enables more free-form natural language instructions from users' input but also enriches the scene's semantics, which helps the reasoning for frontier-based exploration and improves performance (+5.4%). These efforts reduce failure cases of all categories.

5 Navigation in the Real World

We further conduct real robot demonstrations to show our method's ability to understand free-form natural language instructions and perform open-set zero-shot navigation in the real world.

5.1 Real Robot Setup

For robots, we use a TurtleBot3 Burger robot with scalable structures to navigate on the ground. We limit its action space to {`stop`, `move_forward`, `turn_left`, `turn_right`}. As in the simulation, we set the forward distance to 0.25m and the rotation angle to 30 degrees. For robotic perception, we use an Intel RealSense SR300 depth camera to capture RGBD images.

For real-world environments, we select an office with sufficient space and various objects for the robot to navigate. The office contains not only common objects like "chair", "desk", "computer" and "cabinet", but also less common ones like "robot arm" and "3D printer".

5.2 Qualitative Studies

We conduct qualitative studies on our OpenFMNav in the real world, as shown in Fig. 4. The results show that our method can perform effective zero-shot navigation in the real world given free-form natural language instructions. Especially, our method is robust to distractors, open-set objects and free-form demands.

For distractors, rather than object categories, our proposed ProposeLLM can extract the attributes in the instruction ("red chair"), which can be further detected and segmented by PerceptVLM. In Fig. 4(a), we can see that, among the three chairs in the observation, only the red chair is masked.

For open-set objects, due to the large-scale training data of foundation models, our method can also navigate to objects that are uncommon and out-of-vocabulary, such as the "robot arm" in Fig. 4(b).

Another intriguing feature of our method is that our method can adaptively add up goals in the navigation. This happens when the instruction is a free-form demand for ambiguous objects. For example, in Fig. 4(c), when the user needs to wash hands, the ProposeLLM first proposed "sink", "soap" and "towel", but they are not necessarily present in the scene. When the agent explores the environment, the DiscoverVLM can actively discover what's new in the environment and reason about whether they can potentially fulfill the user's demand. In this case, a "tap" is discovered and identified as a goal so that the agent can directly navigate to it without further exploration. This is extremely helpful when the humans are also unaware of the scene details.

6 Conclusions

In this paper, we presented a novel framework, OpenFMNav, for open-set zero-shot object navigation. By leveraging foundation models, our method could understand free-form natural language instructions, conduct reasoning, and perform effective zero-shot object navigation. Extensive experiments showed the superiority of our framework. Finally, we conducted real robot demonstrations to validate our method's open-set-ness and generalizability to real-world environments.

Ethics Statement

In this paper, we present a method for open-set zero-shot object navigation. This method can be used for zero-shot robotic navigation in diverse scenarios, such as home robots, warehouse robots, and so on. Our work further addresses the issue of ambiguous or free-form natural language instructions, benefitting the interaction between humans and robots. However, foundation models can have safety issues and risks such as privacy leaks and jailbreaking (Deng et al., 2023; Chao et al., 2023), which need to be further addressed.

Limitations

While extensive experiments validate the effectiveness of our method design, there exist a number of limitations in our work. First, our method requires relatively accurate depth sensors to build the 2D map, while the observed depths may have much noise in reality, causing performance degradation. Moreover, we acknowledge that our method requires stable Internet connections to get responses from APIs of foundation models, limiting the potential of large-scale deployment in harsh environments. We hope future works on depth sensing, LLM quantization, and edge computing can mitigate such limitations.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. 2020. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models.

Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. 2023. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. *arXiv preprint arXiv:2309.10309*.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries.

Devendra Singh Chaplot, Helen Jiang, Saurabh Gupta, and Abhinav Gupta. 2020a. Semantic curiosity for active visual learning. In *Computer Vision – ECCV 2020*, pages 309–326, Cham. Springer International Publishing.

Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. 2020b. Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12875–12884.

Shizhe Chen, Thomas Chabal, Ivan Laptev, and Cordelia Schmid. 2023. Object goal navigation with recursive implicit maps. In *The 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023)*.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and

Yang Liu. 2023. Masterkey: Automated jailbreak across multiple large language model chatbots.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Vishnu Sashank Dorbala, James F Mullen Jr, and Dinesh Manocha. 2023. Can an embodied agent find your" cat-shaped mug"? llm-based zero-shot object navigation. *arXiv preprint arXiv:2303.03480*.

Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. 2023. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23171–23181.

Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. 2023. Navigating to objects in the real world. *Science Robotics*, 8(79):eadf6991.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. 2022. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7606–7623, Dublin, Ireland. Association for Computational Linguistics.

Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. 2018. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, Online. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded language-image pre-training. In *CVPR*.

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

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. 2022. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. In *Advances in Neural Information Processing Systems*, volume 35, pages 32340–32352. Curran Associates, Inc.

Arjun Majumdar, Fei Xia, brian ichter, Dhruv Batra, and Leonidas Guibas. 2023. Findthis: Language-driven object disambiguation in indoor environments. In *7th Annual Conference on Robot Learning*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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

Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. 2021. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai.

Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. 2022. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale.

Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. 2018. Semi-parametric topological memory for navigation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana

Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347.

James A Sethian. 1999. Fast marching methods. *SIAM review*, 41(2):199–235.

Dhruv Shah, Michael Equi, Blazej Osinski, Fei Xia, Brian Ichter, and Sergey Levine. 2023. Navigation with large language models: Semantic guesswork as a heuristic for planning. *arXiv preprint arXiv:2310.10103*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Hongcheng Wang, Andy Guan Hong Chen, Xiaoqi Li, Mingdong Wu, and Hao Dong. 2023. Find what you want: Learning demand-conditioned object attribute space for demand-driven navigation.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Erik Wijmans, Abhishek Kadian, Ari S. Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. 2019. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *International Conference on Learning Representations*.

Fei Xia, Amir Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. 2018. Gibson env: Real-world perception for embodied agents.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding.

Karmesh Yadav, Santhosh Kumar Ramakrishnan, John Turner, Aaron Gokaslan, Oleksandr Maksymets, Rishabh Jain, Ram Ramrakhya, Angel X Chang, Alexander Clegg, Manolis Savva, Eric Undersander, Devendra Singh Chaplot, and Dhruv Batra. 2022a. Habitat challenge 2022. `https://aihabitat.org/challenge/2022/`.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. 2022b. Habitat-matterport 3d semantics dataset. *arXiv preprint arXiv:2210.05633*.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, Zhou Zhao, Shinji Watanabe, and Helen Meng. 2023a. Uniaudio: An audio foundation model toward universal audio generation.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023b. The dawn of lmms: Preliminary explorations with gpt-4v(ision).

Naoki Harrison Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. 2023. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2nd Workshop on Language and Robot Learning: Language as Grounding*.

Bangguo Yu, Hamidreza Kasaei, and Ming Cao. 2023. L3mvn: Leveraging large language models for visual target navigation. *arXiv preprint arXiv:2304.05501*.

Jiazhao Zhang, Liu Dai, Fanpeng Meng, Qingnan Fan, Xuelin Chen, Kai Xu, and He Wang. 2023. 3d-aware object goal navigation via simultaneous exploration and identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6672–6682.

Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. 2023. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. *arXiv preprint arXiv:2301.13166*.

A API Usage

Model Name	API Name
ProposeLLM	`gpt-4-1106-preview`
DiscoverVLM	`gpt-4-vision-preview`
ReasonLLM	`gpt-4-1106-preview`

Table 3: API usage

B Hyperparameters

Parameter	Value
Discovery Frequency σ_{freq}	0.01
Frontier Goal Update Interval δ	20
Confidence Score Threshold	0.55
LLM/VLM Temperature	0
Initial Prior Objects O_{pri}	See Fig. 5

Table 4: Hyperparameters

chair, bed, plant, toilet, tv, couch, desk, refrigerator, sink, bathtub, shower, towel, painting, trashcan, stairs

Figure 5: Initial prior objects O_{pri}

C Prompts

C.1 Prompts for ProposeLLM

The prompts for ReasonLLM are shown in Fig. 6.

```
SYSTEM_PROMPT="""You are an intelligent embodied agent called ProposeLLM that follows an instruction to navigate in a real indoor environment. Your goal is to propose a list of objects that can satisfy the user's need.

You are firstly given an instruction that indicates the user's need.

If the instruction contains a specific goal object, like "go to the bed" or "find the red bottle", you should directly output the goal object with its possible attributes, like "bed" or "red bottle".

Otherwise, if the instruction is more general, like "I'm so thirsty", you should inference via common sense which objects are feasible and output a list of objects that can satisfy the user's need, like "bottle", "cup", "refrigerator", etc.

Notice that your output should be a list of objects with their possible attributes, even if there is only one object in the list."""

USER1="""go to the bed"""

ASSISTANT1="""Thought: The instruction contains a specific object goal, so I will directly output "bed".

Answer: ["bed"]"""

USER2="""I have been standing for hours. I need some place to sit down and rest."""

ASSISTANT2="""Thought: The instruction is quite general, so I will use my common sense. The user needs some place to sit down, so candidate objects in an indoor scene can be chair, couch, etc. I will output a list of these objects.

Answer: ["chair", "couch"]"""
```

Figure 6: Prompts for ProposeLLM

C.2 Prompts for DiscoverVLM

The prompts for ReasonLLM are shown in Fig. 7.

```
SYSTEM_PROMPT="""You are an intelligent assistant called DiscoverVLM that can understand natural language and scene images. Given a list of objects and an image, your goal is to discover new objects in the image that are not in the list.

You should consider the following rules when discovering new objects:

(1) You should first consider, what's in the image? Note that you should only include objects in the house, and avoid things that are part of the house, like ceiling, wall, floor, window etc and avoid room names, like bedroom, kitchen, etc.

(2) Considering the given object list, you should only output things that are not in the list or are not similar to things in the list because your duty is to discover new things. For example, if the given object list contains "couch" or "tv", you should not output "sofa" or "television" because they are similar.

(3) Confirm that the objects you output are in the image. For example, if the image is a bedroom, you should not output "bathtub" because it is impossible to find a bathtub in a bedroom. And also confirm the objects you output don't violate rule (1).

(4) Avoid objects are common everywhere. For example, objects like light switch and door are present in every room, so you should not output them.

Your output should be in the form of "Answer: <list of objects>" such as:

Answer: ["chair", "bed", "bottle"]
"""

USER="""What objects can you see in the image?"""
```

Figure 7: Prompts for DiscoverVLM

C.3 Prompts for PerceptVLM

For PerceptVLM, given the current object list, we use dots to separate each object as the object prompt p_t.

For example, if the object list is `[chair, bed, plant, toilet, tv, couch]`, the object prompt is "`chair.bed.plant.toilet.tv.couch`".

C.4 Prompts for ReasonLLM

The prompts for ReasonLLM are shown in Fig. 8.

```
SYSTEM_PROMPT="""You are an intelligent embodied agent called ReasonLLM that follows an instruction to navigate in an indoor environment. You are firstly given
an object goal class for you to find, which is called the goal.

Then, at each step, your task is to take several descriptions of what an area contains to output scores for these areas to contain the goal. Each score is a
floating point number between 0 and 1.

Your output should be a list of scores.

At each step, you should consider:

(1) For each description, according to what the area contains, is it possible that the goal is also in this area? To better do reasoning, you can imagine what
kind of room the area is in, for example, a bedroom, a living room, a bathroom, etc. Based on the common sense, you can judge the possibility that the goal to
be in this area.

(2) If the goal class is already in the description, the score should be 1 without any hesitation.

(3) If one area contains nothing, it is still possible that the goal is in that area. Give a score of 0.4 to 0.6. Score the area higher in that case when other
areas are not likely to contain the goal.

(4) If there are no current frontiers, skip the thought and output 'No frontiers'."""

USER1="""Goal: toilet

- Description 0: The area contains a towel, a bathtub and a sink.

- Description 1: The area contains a bed and a plant.

- Description 2: The area contains a sofa, a TV and a table.

"""

ASSISTANT1="""Thought: Let's analyze each description.

- Description 0: this area contains a bathtub and a sink, so it is possibly a bathroom, and the goal is toilet, so it is possible that the goal is in this area,
I will give a score of 0.9

- Description 1: this area contains a bed and a plant, so it is possibly a bedroom, and the goal is toilet, so it is not likely that the goal is in this area. I
will give it a 0.2

- Description 2: this area contains a sofa, a TV and a table, so it is possibly a living room, and the goal is toilet, so it is also not likely that the goal is
in this area. I will give it a 0.3

Answer: [0.9, 0.2, 0.3]"""

USER2="""Goal: bed

- Description 0: The area contains a towel, a bathtub and a sink.

- Description 1: The area contains a bed and a plant.

- Description 2: The area contains a sofa, a TV and a table.

"""

ASSISTANT2="""Thought: Let's analyze each description.

- Description 0: this area contains a bathtub and a sink, so it is possibly a bathroom, and the goal is bed, so it is not possible that the goal is in this
area. My score is 0.1

- Description 1: this area contains a bed, which is the goal, so the score is 1

- Description 2: this area contains a sofa, a TV and a table, so it is possibly a living room. The goal is a bed, so it can be near this area. I will give it a
0.5

Answer: [0.1, 1, 0.5]"""
```

Figure 8: Prompts for ReasonLLM