OVExp: Open Vocabulary Exploration for Object-Oriented Navigation

Meng Wei [1,2] **Tai Wang** [2] **Yilun Chen** [2] **Hanqing Wang** [2]
Jiangmiao Pang [2] ✉ **Xihui Liu** [1] ✉
[1]The University of Hong Kong [2]Shanghai AI Laboratory

Project Page: `https://ovexp.github.io/`

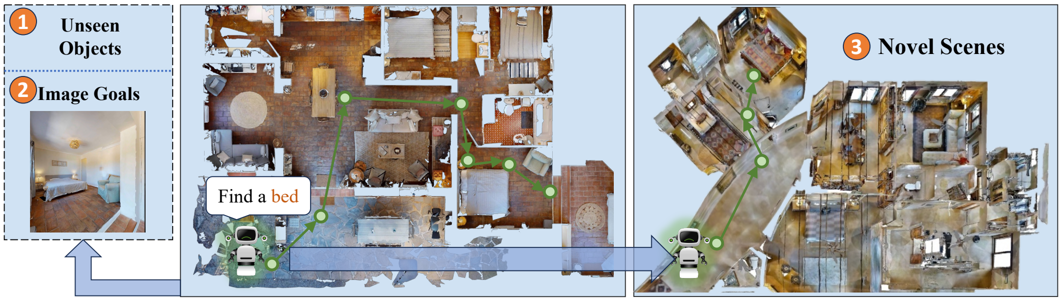

Figure 1: Our learning-based navigation framework enables Open Vocabulary Exploration. Trained with object goals, it generalizes effectively to unseen objects, image goals, and novel scenes, demonstrating robust versatility in diverse navigation tasks.

Abstract

Object-oriented embodied navigation aims to locate specific objects, defined by category or depicted in images. Existing methods often struggle to generalize to open vocabulary goals without extensive training data. While recent advances in Vision-Language Models (VLMs) offer a promising solution by extending object recognition beyond predefined categories, efficient goal-oriented exploration becomes more challenging in an open vocabulary setting. We introduce *OVExp*, a learning-based framework that integrates VLMs for Open-Vocabulary Exploration. *OVExp* constructs scene representations by encoding observations with VLMs and projecting them onto top-down maps for goal-conditioned exploration. Goals are encoded in the same VLM feature space, and a lightweight transformer-based decoder predicts target locations while maintaining versatile representation abilities. To address the impracticality of fusing dense pixel embeddings with full 3D scene reconstruction for training, we propose constructing maps using low-cost semantic categories and transforming them into CLIP's embedding space via the text encoder. The simple but effective design of *OVExp* significantly reduces computational costs and demonstrates strong generalization abilities to various navigation settings. Experiments on established benchmarks show *OVExp* outperforms previous zero-shot methods, can generalize to diverse scenes, and handle different goal modalities.

✉Corresponding authors.

Preprint. Under review.

1 Introduction

The task of object-oriented navigation requires the embodied agent to locate specified object goals in unseen environments and navigate to them. The goal can be defined either by language prompts, typically specifying the object's category [2] or by an image depicting the target object [14]. *Accurately identifying the object* while *exploring the environment efficiently* is the core of this problem. Recent advances in Vision-Language Models (VLMs) like CLIP [21] have expanded object recognition beyond predefined categories, and some recent works [10, 11] utilize this ability for goal identification in navigation, combined with the classical frontier-based exploration (FBE) [30] algorithm. However, achieving efficient goal-oriented exploration, which involves reasoning about room layouts, object arrangements, or object relationships, is still an understudied problem.

Earlier works [4, 12, 23, 22, 33] implicitly integrate scene common sense for semantic exploration by reinforcement learning or imitation learning. However, these learning-based methods are limited to a closed-set setting which requires sufficient training data for each goal. Recently, some training-free approaches [35, 31, 26, 7, 24, 9] have emerged to use the rich commonsense knowledge in Large Language Models (LLMs) for planning. They convert each observation image into language input and query LLMs for the next goal, making such methods cumbersome and inefficient. Furthermore, language descriptions extracted from sparse observation images cannot provide a comprehensive context of the 3D physical world, leading to inaccurate LLM decision-making. Therefore, it is critical to integrate VLMs into the framework in a compact way, transferring their powerful generalization capability to enable *scene-aware exploration* for open-vocabulary object goals.

In this paper, we make the first effort to integrate VLMs into a new learning-based framework for Open-Vocabulary Exploration, *OVExp*. We demonstrate that the powerful visual-language representations embedded within VLMs, pretrained on vast corpora of natural images and texts, can be transferred to enable generalizable semantic-map-based exploration. In the same spirit as the construction of VLMaps [11], *OVExp* takes ego-centric RGB-D images as input, encodes their visual-language features with VLMs, and projects these features onto top-down maps.The object goal, in text or image form, is encoded in the same feature space as the maps. At each time step, the map is updated with new observation features. The updated map is then downsampled and flattened into a sequence of patch tokens, which are fused with the goal embedding for goal-conditioned exploration. Finally, a lightweight transformer-based decoder is employed to predict the target location in the map for subsequent local planning. However, constructing maps using pixel-level features from VLMs like LSeg [16] will cause high computational costs and storage demands, making it impractical to train such an exploration policy network.

To address this problem, we propose a novel strategy to achieve a trade-off between map scalability and training cost to train the network efficiently. Specifically, leveraging CLIP's joint visual-language space, we construct the maps by first creating low-cost semantic categorical maps and then transforming them into language feature-based maps through CLIP's text encoder. During inference, the framework seamlessly transitions to vision-based mapping to handle open vocabulary object goals. This transferable vision and language mapping strategy can significantly reduce the computational cost. Surprisingly, despite trained on a limited set of object categories, *OVExp* demonstrates robust generalization capabilities. We evaluate this framework on established object-oriented navigation benchmarks under three open vocabulary settings: 1) *Zero-Shot Setting*: *OVExp* surpass previous zero-shot methods by a large margin on the HM3D Object Goal Navigation task. 2) *Cross-Dataset Setting*: *OVExp*, trained on HM3D, shows impressive generalization ability when transferring to MP3D, approaching in-domain supervised methods without further fine-tuning. 3) *Cross-Modality Setting*: *OVExp*, trained with textual goals, can effectively handling different modalities of open-vocabulary goals on HM3D Instance Image Goal Navigation Task. In summary, OVExp offers *a simple but effective* framework for exploration and demonstrates remarkable adaptability in handling diverse open vocabulary navigation scenarios, leveraging CLIP's joint visual-language space.

2 Related Work

Map-based Navigation. Map-based representation has shown great promise in navigation tasks. Constructing the mapping of the environment can enable efficient path planning with spatial awareness and history information. Various types of maps have been proposed, including occupancy maps [5], topological graphs [6, 26], semantic maps [4, 32] and implicit maps [8]. Among them, as the semantic maps can integrate high-level semantic information into the navigation process, follow-up works [32, 36, 22, 33] explicitly predict the probabilities of the long-term goal locations on the map, which have achieved state-of-the-art performance without expensive reinforcement learning. Moreover, leveraging advancements in powerful semantic representation learning facilitated by Large Vision-Language Models (VLMs), VLMaps [11] extends traditional categorical semantic maps to high-dimensional semantic maps by projecting pixel features from LSeg [16] onto the top-down map. However, VLMaps necessitates the construction of a complete map initially, followed by open vocabulary goal indexing. In contrast, we focus on improving the exploration efficiency and investigate how VLMs can empower open vocabulary goal-oriented exploration.

Open Vocabulary Object Navigation. Navigating towards open vocabulary object goals is a more realistic scenario, and numerous researchers have recently delved into this challenging problem. Several benchmarks [10, 17, 13] have been designed for open vocabulary navigation tasks to promote the development of solutions. Some previous methods [18] adopt images as goals, which are encoded with CLIP to generalize to diverse objectives. A recent trend involves leveraging Large Language Models (LLMs) for training-free exploration approaches [35, 31, 26, 7, 24]. These methods exploit the commonsense knowledge embedded in LLMs to interactively explore a range of goals. However, these methods rely solely on observation images, potentially limiting their ability to grasp a holistic understanding of the 3D environment and leading to suboptimal decision-making. Hence, there still remains a gap in developing exploration algorithms capable of effectively navigating through the open world, including generalizing to unseen goals, multi-modal goals, and diverse scenes.

3 Method

3.1 Preliminary: Object-Oriented Navigation Task Definitions

We target an open vocabulary setting for object-oriented navigation. Instead of making specific solutions for individual tasks, our objective is to tackle the following two object-oriented navigation tasks in a unified manner. We believe that handling goals specified across different modalities is a crucial aspect of an open vocabulary navigation system.

Object Goal Navigation **(ObjectNav)** requires an embodied agent to navigate to an instance of a specified object category, like "chair" or "table", in unseen environments, using RGB and Depth cameras and a 3-DoF current pose. The agent explores its surroundings until it locates the first instance of the target object category. While in Instance-Specific Image Goal Navigation **(InstanceImageNav)**, the agent is required to navigate to a particular object instance depicted in a provided RGB image, regardless of its perspective. This task is more complex as it requires the agent to precisely identify one instance. Both tasks are object-centric, with the same evaluation protocol: episode ends upon execution of the stop action or upon reaching the timestep limit. Success is achieved when the agent predicts a stop action at a location where the distance to the target object is less than 1 meter and the target object is within view. The discrete action space $a_t \in \mathcal{A}$ includes `move_forward`, `turn_left`, `turn_right`, `look_up`, `look_down`, and `stop`.

3.2 Overview of the Modular Framework

We propose a modular framework especially for achieving Open Vocabulary Exploration (*OVExp*). As shown in Figure 2, the proposed *OVExp* framework mainly consists of three modules: (1) The *Transferable Vision and Language Mapping* module for constructing a visual/language memory map to comprehend the scene holistically. (2) The *Goal-Conditioned Exploration Policy* module for learning scene priors from the map representations and predicting the long-term goal location with a goal-conditioned exploration network. (3) The *Local Policy* module for analytical planning which determines the next waypoint towards the predicted long-term goal.

3.3 Transferable Vision and Language Mapping

We adopted different types of high-dimensional semantic maps during training and testing phrases, because collecting semantic maps with dense pixel-level visual embeddings from LSeg [16] is memory-intensive for training. Therefore, we employ the semantic mapping procedure detailed in Section 3.3.1 to construct categorical semantic maps. We curate a list of 92 object categories from the

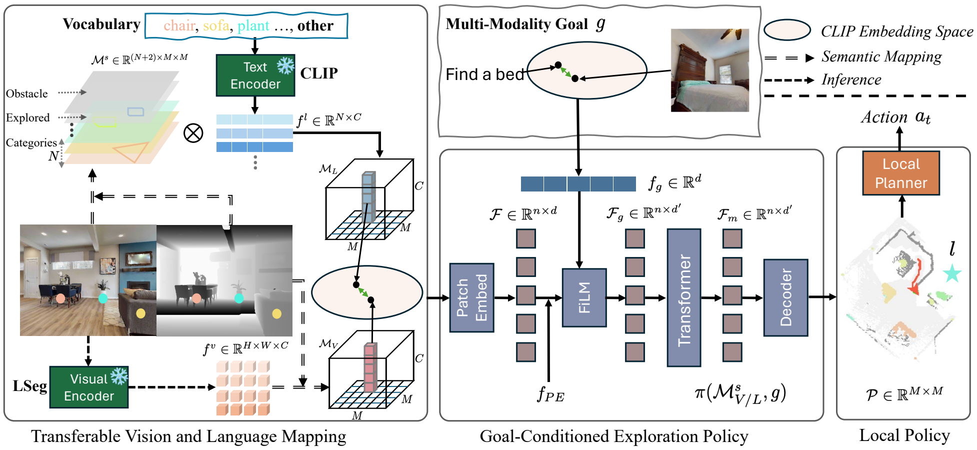
Vocabulary
chair, sofa, plant ..., other
Text Encoder
CLIP
$\mathcal{M}^s \in \mathbb{R}^{(N+2)\times M\times M}$
Obstacle
Explored
Categories
N
$f^l \in \mathbb{R}^{N\times C}$
$\mathcal{M}_L$
C
M
M
$\mathcal{M}_V$
C
M
M
$f^v \in \mathbb{R}^{H\times W\times C}$
LSeg
Visual Encoder
Transferable Vision and Language Mapping
Multi-Modality Goal g
Find a bed
CLIP Embedding Space
Semantic Mapping
Inference
Action a_t
$f_g \in \mathbb{R}^d$
$\mathcal{F} \in \mathbb{R}^{n\times d}$
$\mathcal{F}_g \in \mathbb{R}^{n\times d'}$
$\mathcal{F}_m \in \mathbb{R}^{n\times d'}$
Patch Embed
FiLM
Transformer
Decoder
f_{PE}
$\pi(\mathcal{M}^s_{V/L}, g)$
Goal-Conditioned Exploration Policy
Local Planner
l
$\mathcal{P} \in \mathbb{R}^{M\times M}$
Local Policy

Figure 2: The overall framework of OVExp for open vocabulary object-oriented exploration. OVExp can accept either language-based or vision-based maps as input and accommodates textual and visual object goals. For simplicity, the goal identification model is omitted.

HM3DSem [29] dataset, which offers abundant pixel-level annotated semantics. These categorical maps are subsequently transformed into language-based maps using CLIP's text embeddings, detailed in section 3.3.2. While in inference, we directly construct maps through vision-based mapping 3.3.3, as used in VLMaps [11].

3.3.1 Semantic Mapping

In learning-based goal-oriented exploration, semantic maps [4] have proven to be an effective representation for encoding episodic navigation history. Semantic maps can capture not only geometric priors like spatial layout, obstacles, and navigable areas but also semantic information such as scene object categories, their locations, and spatial relationships.

To construct the semantic map, egocentric visual observation is first segmented into semantic categories using a pre-trained segmentation model. Next, a point cloud is extracted from the depth image by back-projection into the 3D world. Each point is associated with the corresponding semantic label in the 2D observation image. The point cloud is then binned into a voxel grid. Summing along the height dimension, the voxel grid is projected into an egocentric top-down semantic map. To be merged with the global map, the egocentric map is finally transformed into the allocentric coordinate system using agent's pose information. The constructed semantic map $\mathcal{M}^s \in \mathbb{R}^{(N+2)\times M\times M}$ has 2 non-semantic channels (1: obstacle map, 2: explored map) and N semantic channels related to N object categories of interests. Each cell within the $M \times M$ grid corresponds to a region of $5cm \times 5cm$.

However, categorical semantic maps face scalability challenges when applied to a large vocabulary. Hence, to address the open vocabulary setting, well-pretrained representations become essential.

3.3.2 Language-based Mapping

To enrich the representation of collected semantic maps, we encode the N semantic channels of $\mathcal{M}^s$ with language features derived from CLIP's embedding space. Formally, we input the text list of N objects $\{O_1, O_2, \ldots, O_N\}$ (O_N represents the background class and use "other" as the text) into CLIP's text encoder to produce high-dimensional language-based object features $f^l \in \mathbb{R}^{N\times C}$. Each semantic channel $\mathcal{M}^s_{O_i} \in \mathbb{R}^{1\times M\times M}$ of the semantic map encodes the confidence score of the object's existence in the grid cells. To transform $\mathcal{M}$ into a language-enhanced map representation $\mathcal{M}_L$, we compute the weighted sum of the semantic channels and the corresponding language embeddings:

$$\mathcal{M}_L = \frac{\sum_{i=1}^{N} \mathcal{M}^s_{O_i} * f^l_{O_i}}{\sum_{i=1}^{N} \mathcal{M}^s_{O_i}} \tag{1}$$

where $\mathcal{M}_L \in \mathbb{R}^{C\times M\times M}$. In $\mathcal{M}_L$, the objects are no longer encoded in separate channels; instead, each grid cell contains the averaged language features of all objects present.

3.3.3 Vision-based Mapping

During inference, we construct an equivalent vision-based map input with the pretrained image encoder of LSeg, which produces pixel embeddings aligned with their corresponding label embeddings, resembling real-world scene representations. The vision-based map construction is a one-stage process. At each timestep t, we extract the dense pixel-wise visual features $f^v \in \mathbb{R}^{H\times W\times C}$ of the observation RGB image $\mathcal{I} \in \mathbb{R}^{H\times W}$. The transformation from 2D egocentric pixel representations to a top-down grid map is similar to the semantic mapping process described in Section 3.3.1. The only difference lies in how the grid map is updated during the online mapping process. In updating the semantic categorical maps, the maximum object confidence from all timesteps is consistently taken for each grid cell. However, the vision-based map $\mathcal{M}_V^t$ is updated as follows:

$$\mathcal{M}_V^t\left[i,j\right] = \frac{\mathcal{M}_V^{t-1}\left[i,j\right] \times \mathcal{N}^{t-1}\left[i,j\right] + m_v^t\left[i,j\right]}{\mathcal{N}^{t-1}\left[i,j\right]+1}; \mathcal{N}^t\left[i,j\right] = \mathcal{N}^{t-1}\left[i,j\right]+1 \tag{2}$$

where $\mathcal{M}_V \in \mathbb{R}^{C\times M\times M}$. $[i,j]$ represents the locations that will be updated by the incoming map feature $m_v^t \in \mathbb{R}^{C\times M\times M}$ which is projected from the current observation feature f^v. $\mathcal{N} \in \mathbb{R}^{M\times M}$ records the number of updates in each grid cell over time.

3.4 Goal-Conditioned Exploration Policy

Given robust high-dimensional semantic maps $\mathcal{M}^s_{V/L} \in \mathbb{R}^{(C+2)\times M\times M}$ (the first two channels represent the obstacle map and explored area), which encode the spatial and semantic information of the environment, and an object goal g, our objective is to learn a global policy π that outputs the long-term goal location l within the local map:

$$\pi(\mathcal{M}^s_{V/L}, g) \rightarrow l \tag{3}$$

We use $\mathcal{M}^s_L$ for training and switch to $\mathcal{M}^f{}_V$ during inference. Following previous map-based methods [32, 33], the policy π will output an object probability map and the long-term goal location l is selected with the largest probability. This policy is flexible to handle goals specified in different modalities, such as a textual goal g_t in ObjectNav and an image goal g_i in InstanceImageNav.

Goal-Conditioned Map Encoder. The input map $\mathcal{M}^s_{V/L}$ is first partitioned into non-overlapping patches which are projected into map token embeddings through the patch embed operation. Then we add a learnable positional embedding f_{PE} to the token embeddings and obtain the map features $\mathcal{F}$:

$$\mathcal{F} = \textsc{PatchEmbed}(\mathcal{M}^s_{V/L}) + f_{PE} \tag{4}$$

where $\mathcal{F} \in \mathbb{R}^{n\times d}$, n denotes the number of map tokens and d denotes the hidden size.

To condition the long-term goal prediction on the object goal embedding $f_g \in \mathbb{R}^d$, which is text or image embedding from CLIP, we employ an efficient Feature-wise Linear Modulation (FiLM) [20] layer. This layer applies a feature-wise affine transformation to fuse the two sources of features $\mathcal{F}$ and f_g , producing a goal-conditioned map feature $\mathcal{F}_g$:

$$\mathcal{F}_g = \gamma(f_g) \odot h(\mathcal{F}) + \beta(f_g); \mathcal{F}_m = \textsc{Transformer}(\mathcal{F}_g) \tag{5}$$

where $\mathcal{F}_g \in \mathbb{R}^{n\times d'}$.$\gamma(.)$, $\beta(.)$ and $h(.)$ are three linear transformations. $\gamma(.)$ and $\beta(.)$ generate the scaling and shifting vectors from f_g respectively. $h(.)$ reduces the dimension of $\mathcal{F}$ from $d = 512$ to $d' = 64$. This fusion process effectively integrates the semantic goal information into the map features. Finally, the encoded map features $\mathcal{F}_m \in \mathbb{R}^{n\times d'}$ is produced by feeding $\mathcal{F}_g$ to a two-layer transformer to facilitate feature representation learning.

Goal Location Prediction. To generate the object probability map for selecting the goal location, we design a convolution network as the decoder. The decoder $\mathcal{D}$ consists of one convolution layer and two transposed convolution layers which upsamples the map feature $\mathcal{F}_m$ to the original map resolution and generate the goal probability map $\mathcal{P} \in \mathbb{R}^{M\times M}$. The location with the highest value in this map is selected as the predicted long-term goal location. Following [33], to improve the efficiency of exploration, the final long-term goal location is determined by weighting $\mathcal{P}$ with the geodesic distance to the agent's current location:

$$\mathcal{P} = \mathcal{D}(f_g); l = arg\max_{i,j}(\mathcal{P}_{ij} \times g_{ij}) \tag{6}$$

(i, j) denotes the map index and g is the exponential weight derived from the geodesic distance.

We use the binary cross-entropy loss to train the goal-conditioned exploration policy. As training with the full list of objects results in substantial computational cost, we employ a federated loss. At each training step, a subset of objects is selected to compute the binary cross-entropy loss. This subset is dynamically chosen to balance computational efficiency and model robustness.

3.5 Analytical Local Planner

To reach the predicted goal locations from exploration policy or the identified goal points, we adopted an analytical local planner to translate the long-term goal location into an executable action $a_t \in \mathcal{A}$ as in previous modular map-based methods. The Fast Marching Method (FMM) is employed incrementally to calculate the shorted path from the agent's current location to the goal location. Then a waypoint along this path is selected considering the agent's step distance. As the agent sometimes gets stuck in obstacles, we adopted an untrap strategy to help the agent to recover from the stuck.

4 Experiment

4.1 Experimental Setup

Training Dataset. We generate a dataset of semantic maps collected from the HM3DSem [29] dataset using the Habitat [25] simulator. HM3DSem annotates object instances across 216 3D scene reconstructions with 1660 raw object names. We follow the official splits of the 2022 Habitat ObjectNav Challenge [27], utilizing $80/20/20$ scenes for training, validation, and test sets, respectively. Due to the significant noise in HM3DSem's raw annotations and the impracticality of building semantic maps from such a large list of objects, we curated a list of 92 objects from HM3DSem. Next, we set a goal-agnostic agent to explore the training scenes from random start locations, allowing it to wander for 500 steps. During this process, we save the semantic map every 25 steps. To avoid using maps with minimal unexplored areas, we select only the first half of the saved maps for training. A total of 27983 partial semantic maps are collected to train the exploration policy.

Evaluation Datasets. The evaluation is conducted on the validation sets of three object-centric navigation datasets, including one InstanceImageNav dataset and two ObjectNav datasets:
`HM3D-ObjectNav` [27]: Released in the Habitat 2022 challenge, this dataset comprises 2000 episodes from 20 validation scenes in HM3D, targeting 6 specific goal objects.
`MP3D-ObjectNav` [1]: Released in the Habitat 2020 challenge, this dataset comprises 2195 episodes from 11 validation scenes in MP3D, targeting 21 specific goal objects.
`HM3D-InstanceImageNav` [28]: Released in the Habitat 2023 challenge, this dataset comprises 1000 episodes from 36 validation scenes in HM3D, targeting the goal images of 6 specific objects.

Evaluation Metrics. We evaluate navigation performance using two standard metrics. ***Success***: measures the proportion of episodes in which the agent successfully stops near the goal object location. ***SPL*** (Success weighted by Path Length): assesses the efficiency of the navigation by weighting the success rate by agent path length relative to the oracle shortest path length.

4.2 Implementation Details

We use a global map of size 960×960 for training, applying random crop operations to 720×720, along with random flips and random rotations for data augmentation. We embed the map features with a patch size of 16×16. Then a 2-layer transformer with the hidden size of 512 and 8 attention heads is used to update the map features. For fusion with FiLM, both the encoded map features and goal embeddings are reduced to a hidden size of 64. For the transposed convolutional decoder, the first convolutional layer uses a kernel size of 3 and a padding size of 1. The two transposed convolutional layers upsample the feature maps using transposed kernels with a size of 4. The prediction model is trained for 20 epochs with a batch size of 8, requiring 20 hours with 8 NVIDIA V100 GPUs. We use the AdamW optimizer with the initial learning rate of 1^{e-4} and weight decay of 1^{e-4}. We use cosine decay for learning rate decay.

Table 1: Comparison with state-of-the-art ObjectNav methods on the validation set of HM3D. *indicates results run from their officially released checkpoint.“_” means the second best results.

Setting	Method	Semantic Map	Training	LLM Planner	Success↑	SPL↑
Supervised	SemExp [4]	✓	RL	-	37.9	18.8
	PIRLNav [23]	×	RL & IL	-	**61.9**	27.9
	PEANUT* [33]	✓	SL	-	60.5	**30.7**
	OVExp	×	SL	-	60.6	29.7
Zero-Shot	CoW [10]	✓	×	×	32.0	18.1
	ZSON [18]	✓	✓	×	25.5	12.6
	L3MVN [31]	✓	×	GPT-2	50.4	23.1
	PixelNav [3]	×	✓	GPT-4	37.9	20.5
	ZSC [35]	✓	×	GPT-3.5	39.2	22.3
	VoroNav [26]	✓	×	GPT-3.5	42.0	26.0
	OVExp-ZS	✓	✓	×	**59.7**	**28.8**

Table 2: Cross-Dataset ObjectNav Performance on the validation set of MP3D.

Method	Train Set	Success↑	SPL↑
THDA [19]	MP3D	28.4	11.0
PONI [22]	MP3D	**31.8**	**12.1**
ZSON [18]	HM3D	15.3	4.8
OVExp	HM3D	28.0	12.0

Table 3: Cross-Modality Performance on the validation set of HM3D-InstanceImageNav. Agent configurations: Stretch and LoCoBot.

Method	Config	Success↑	SPL↑
Mod-IIN [15]	Stretch	56.1	**23.3**
OVExp	Stretch	59.7	21.5
OVExp	LoCoBot	**63.0**	20.5

4.3 Zero-Shot Object Navigation Performance.

Settings. We evaluate the zero-shot generalization ability of our exploration policy by testing our OVExp agent's performance in navigating to novel goal objects. We conduct this zero-shot experiment on the evaluation episodes of HM3D-ObjectNav, which are generated for 6 specific objects. To ensure our exploration policy acquires zero experience with these objects, we train a zero-shot OVExp policy (OVExp-ZS) by excluding these objects, i.e., their locations are not learned during training.

Baselines. We compare our zero-shot agent with three types of existing zero-shot navigation methods categorized by the exploration policy:

- `Heuristic-based Exploration`: CoW [10] combines a goal-agnostic Frontier-Based Exploration (FBE) algorithm with an open vocabulary goal detector.
- `Learning-based Exploration`: ZSON [18] also introduces a learning-based approach for open-world object-goal navigation. By training the policy on extensive image-goal navigation episode data, the agent can be directly transferred to object-goal navigation by embedding both types of goals in the same CLIP embedding space.
- `LLM-based Exploration`: Most existing zero-shot methods are training-free and rely on Large Language Models (LLMs) for decision-making. L3MVN [31] and ESC [35] select one frontier point as the long-term goal, identified through interactions with LLMs. PixelNav [3] employs LLMs to choose the next goal place from panoramic images and then navigates to the identified pixel goal in the selected image with a transformer-based navigation policy. VoroNav [26] generates a reduced voronoi graph on the semantic map and selects a graph node as the long-term goal combining the topological information and LLMs' suggestions.

Moreover, we provide some state-of-the-art closed-set object navigation baselines for better reference on the benchmark: SemExp [4] trains a semantic exploration policy based on semantic maps. PEANUT [33] trains a goal location prediction model to determine the long-term goal in unexplored areas. PIRLNav [23] firstly pretrains the policy with behaviour cloning and then finetunes it with reinforcement learning. We also report the performance of **OVExp** trained on all 92 objects.

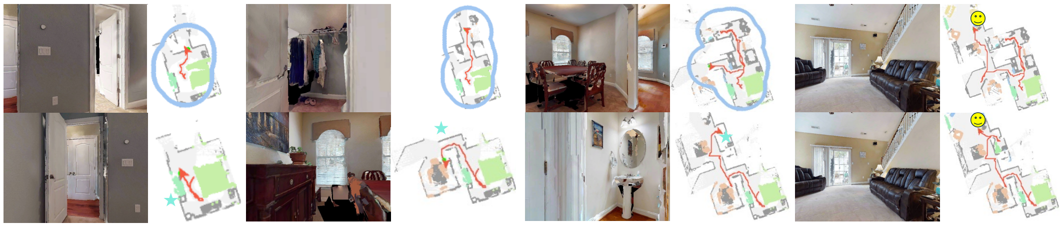

Figure 3: Qualitative results of Zero-Shot object navigation on `HM3D-ObjectNav`. First row: The navigation trajectory of *FBE*. Second row: The navigation trajectory of *OVExp-ZS*.

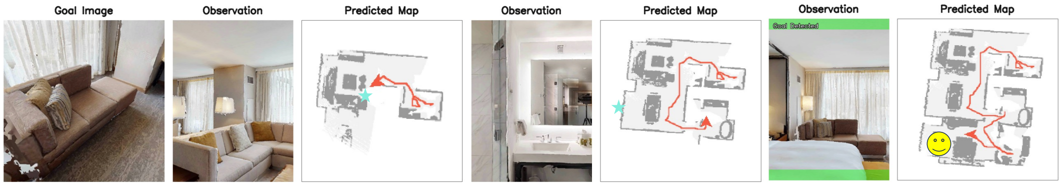

Figure 4: Qualitative results of Cross-Modality object navigation on `HM3D-InstanceImageNav`.

Performance Analysis. As shown in Table 1, **OVExp-ZS** achieves state-of-the-art performance compared with existing learning-based and LLM-based zero-shot ObjectNav methods, with a $+\%$ improvement in success rate and a $+\%$ increase in SPL on HM3D. In the closed-set setting, **OVExp** demonstrates comparable performance to PEANUT. While PEANUT employs PSPNet [34] as the prediction model which is trained on semantic maps constructed from 10 semantic categories, **OVExp** uses a more light-weight model architecture capable of predicting a broader range of object goals.

4.4 Cross-Dataset Performance.

Settings. To further evaluate its generalization ability, we directly test **OVExp**, which is trained on the HM3D dataset, on the MP3D dataset. As MP3D features larger house sizes and more complex scene layouts than HM3D, this cross-dataset evaluation can measure how well the learned open vocabulary exploration policy adapts to environments with diverse scene layouts and object configurations.

Baselines. We mainly compare with ZSON [18] because it is trained on HM3D Image Goal navigation data and tested on MP3D in a similar way. Additionally, we compare with two modular-based methods which are fully trained on MP3D, THDA [19] and PONI [22]. For a fair comparison, we employ the same goal segmentation model with them.

Performance Analysis. As shown in Table 2, **OVExp** demonstrates superior cross-domain generalization ability compared to ZSON [18]. While both **OVExp** and ZSON utilize CLIP embeddings for goal representation, ZSON relies on massive episode data to train an end-to-end RL-based policy. In contrast, our map-based modular framework requires less data and avoids the high computational cost of RL training. Furthermore, **OVExp** achieves very competitive performance with the supervised baselines without training on the challenging MP3D benchmark.

4.5 Cross-Modality Performance.

Settings. We also show that **OVExp** can handle goals across different modalities. To verify this, we evaluate OVExp on the InstanceImageNav task, which is more challenging than ObjectNav because the agent must locate a specific object rather than just the first encountered instance. Hence, the OVExp agent will sequentially navigate to the possible goal locations until the target object is found.

Baselines. The state-of-the-art modular method Mod-IIN [15] proposes a instance re-identification module for goal matching and use FBE as the exploration policy. Since **OVExp** is flexible and can be integrated into any modular framework, we replace FBE in Mod-IIN with **OVExp** for comparison.

Performance Analysis. Our **OVExp** policy is trained on semantic map data collected by an agent configuration that resembles a LoCoBot, whereas Mod-IIN is tested with an embodiment similar to Stretch. Hence, we conduct a cross-embodiment evaluation using Mod-IIN's agent configuration.

Table 4: Performance of using vision and language maps with different sizes during training and inference on HM3D Val split.

Map Size	20 × 20	30 × 30	45 × 45
Success↑	56.8	58.9	**60.3**
SPL↑	26.6	27.8	**29.8**

Table 5: Comparison of using vision-based or language-based maps during inference on HM3D Val split.

Map Type	Success↑	SPL↑
Language	59.6	28.7
Vision	**60.3**	**29.8**

Also, we evaluate **OVExp** using its own agent configuration. The results are reported in Table 3. We observe that **OVExp** exhibits higher success rates but lower performance on SPL compared to Mod-IIN in both embodiments. These findings suggest that our **OVExp** policy, learned for category-based prediction, may not efficiently locate specific instances. However, it still achieves a better search strategy than FBE to find the target object eventually. Furthermore, we verify that **OVExp** has the ability to generalize across different embodiments.

4.6 Ablation Study.

We conduct two ablations to assess the effectiveness of OVExp in achieving a trade-off between computational cost and performance:

Map size. The use of global high-dimensional semantic maps for training incurs the most significant computational cost in our framework. We analyze how different map sizes can impact effectiveness considering our computational resources. We examine patch sizes of $36, 24, 16$, resulting map sizes of $20 \times 20, 30 \times 30, and 45 \times 45$ respectively. As shown in table 4, using maps with higher resolution contributes to better performance. However, further increasing the map size will result in prohibitively computational cost. Therefore, we adopt the map size of 45×45 in this paper.

Impact of Language-to-Vision Map Transfer. To investigate the disparity between the language-based map and the vision-based map during inference, we evaluate the performance using the language-based map with ground truth semantic annotations. The results are reported in Table 5. Surprisingly, using the vision-based maps even outperforms the ground-truth language-based maps. This indicates that the vision-based maps provide a more contextually rich representation of the environment, leading to better generalization to novel scenes and thus improving the performance.

5 Qualitative Results.

As *OVExp* adopts a trade-off between predicted goal locations and closest locations, we conducted a comparison against a pure FBE-based baseline to demonstrate the superior decision-making efficiency of *OVExp* in Zero-Shot object navigation. As shown in Figure 3, while FBE initially explores the coatroom, leading to inefficient back-and-forth movements, *OVExp* directly navigates out of the bedroom and onwards.

Furthermore, we offer a qualitative analysis of transferring *OVExp* to the InstanceImageNav task. As *OVExp* is primarily object-oriented, it may lack a nuanced understanding of specific goal details within an image. However, it still manages to generate a reasonable exploration path. As shown in Figure 4, *OVExp* first generates a goal location in the living room, where another "sofa" is present, before proceeding to the bedroom to locate the goal instance.

More qualitative videos can be viewed on our Project Page.

6 Conclusion

In this paper, we introduced *OVExp*, a novel modular framework that leverages Vision-Language Models (VLMs) for learning-based open-vocabulary exploration. Our approach encodes RGB-D observations with VLMs and project them onto top-down high-dimensional semantic maps, providing a richer contextual representation for generalized object-oriented navigation. Specifically, a novel transferable vision and language mapping strategy is designed to ensure efficient training, relying on the visual-language feature alignment ability of VLMs. Based on this strategy, we train a goal-conditioned exploration policy with language-based maps and change to the vision-based maps during inference. The overall design contributes to a flexible framework which can not only process multi-modality goals but can also generalize to diverse object goals and scenes. Experiments on the zero-shot object goal navigation, cross-dataset object goal navigation and the cross-modality instance-image goal navigation, verify the generalization ability and robustness of *OVExp*.

Limitations and Negative Societal Impacts. While *OVExp* demonstrates generalization across objects, modalities, and scenes, the current version mainly focuses on the object-centric and single-goal navigation tasks. It lacks the ability to handle complex language-based navigation instructions, which may require further assistance from LLMs. We believe that achieving human-like navigation in the complex world necessitates the end-to-end learning with both VLMs and LLMs. The deployment of such navigation systems may also raise several societal concerns. Advanced navigation technologies could be misused for surveillance purposes, potentially infringing on individual privacy. Furthermore, the reliance on extensive visual and language datasets may inadvertently reinforce existing biases present in the data, leading to unfair treatment or discrimination.

References

[1] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. In *arXiv:2006.13171*, 2020.

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

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

[4] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020.

[5] Devendra Singh Chaplot et al. Learning to explore using active neural SLAM. In *ICLR*, 2020.

[6] Devendra Singh Chaplot et al. Neural topological SLAM for visual navigation. In *CVPR*, pages 12875–12884, 2020.

[7] Junting Chen, Guohao Li, Suryansh Kumar, Bernard Ghanem, and Fisher Yu. How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers, 2023.

[8] Shizhe Chen, Thomas Chabal, Ivan Laptev, and Cordelia Schmid. Object goal navigation with recursive implicit maps. In *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[9] Vishnu Sashank Dorbala, James F Mullen Jr, and Dinesh Manocha. Can an embodied agent find your" cat-shaped mug"? llm-guided exploration for zero-shot object navigation. *ICML*, 2023.

[10] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. *CVPR*, 2023.

[11] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[12] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[13] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. *arXiv preprint arXiv:2404.06609*, 2024.

[14] Jacob Krantz, Stefan Lee, Jitendra Malik, Dhruv Batra, and Devendra Singh Chaplot. Instance-specific image goal navigation: Training embodied agents to find object instances. *arXiv preprint arXiv:2211.15876*, 2022.

[15] Jacob Krantz, Theophile Gervet, Karmesh Yadav, Austin Wang, Chris Paxton, Roozbeh Mottaghi, Dhruv Batra, Jitendra Malik, Stefan Lee, and Devendra Singh Chaplot. Navigating to objects specified by images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[16] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *ICLR*, 2022.

[17] Ji Ma, Hongming Dai, Yao Mu, Pengying Wu, Hao Wang, Xiaowei Chi, Yang Fei, Shanghang Zhang, and Chang Liu. Doze: A dataset for open-vocabulary zero-shot object navigation in dynamic environments. *arXiv preprint arXiv:2402.19007*, 2024.

[18] Arjun Majumdar, Gunjan Aggarwal, Bhavika Suresh Devnani, Judy Hoffman, and Dhruv Batra. ZSON: Zero-shot object-goal navigation using multimodal goal embeddings. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=VY1dqOF2RjC`.

[19] Oleksandr Maksymets, Vincent Cartillier, Aaron Gokaslan, Erik Wijmans, Wojciech Galuba, Stefan Lee, and Dhruv Batra. Thda: Treasure hunt data augmentation for semantic navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[20] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[22] Santhosh K. Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Computer Vision and Pattern Recognition (CVPR), 2022 IEEE Conference on*. IEEE, 2022.

[23] Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17896–17906, 2023.

[24] Dhruv Shah, Michael Equi, Blazej Osinski, Fei Xia, Brian Ichter, and Sergey Levine. Navigation with large language models: Semantic guesswork as a heuristic for planning. In *7th Annual Conference on Robot Learning*, 2023.

[25] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *NeurIPS*, 2021.

[26] Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. Voronav: Voronoi-based zero-shot object navigation with large language model, 2024.

[27] 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. Habitat challenge 2022. `https://aihabitat.org/challenge/2022/`, 2022.

[28] Karmesh Yadav, Jacob Krantz, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Jimmy Yang, Austin Wang, John Turner, Aaron Gokaslan, Vincent-Pierre Berges, Roozbeh Mootaghi, Oleksandr Maksymets, Angel X Chang, Manolis Savva, Alexander Clegg, Devendra Singh Chaplot, and Dhruv Batra. Habitat challenge 2023, 2023.

[29] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *ICCV*, 2023.

[30] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, pages 146–151. IEEE, 1997.

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

[32] Zhen Zeng, Adrian Röfer, and Odest Chadwicke Jenkins. Semantic linking maps for active visual object search. *ICRA*, 2020.

[33] Albert J Zhai and Shenlong Wang. PEANUT: Predicting and navigating to unseen targets. In *ICCV*, 2023.

[34] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

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

[36] Minzhao Zhu, Binglei Zhao, and Tao Kong. Navigating to objects in unseen environments by distance prediction. *IROS*, 2022.