Imagine Before Go: Self-Supervised Generative Map for Object Goal Navigation

Sixian Zhang[1,2], Xinyao Yu[1,2], Xinhang Song[1,2], Xiaohan Wang[1], Shuqiang Jiang[1,2,3]
[1]Key Lab of Intelligent Information Processing Laboratory of the Chinese Academy of Sciences (CAS),
Institute of Computing Technology, Beijing, [2]University of Chinese Academy of Sciences, Beijing
[3] Institute of Intelligent Computing Technology, Suzhou, CAS
{sixian.zhang, xinyao.yu, xinhang.song, xiaohan.wang}@vipl.ict.ac.cn, sqjiang@ict.ac.cn

Abstract

The Object Goal navigation (ObjectNav) task requires the agent to navigate to a specified target in an unseen environment. Since the environment layout is unknown, the agent needs to infer the unknown contextual objects from partially observations, thereby deducing the likely location of the target. Previous end-to-end RL methods capture contextual relationships through implicit representations while they lack notion of geometry. Alternatively, modular methods construct local maps for recording the observed geometric structure of unseen environment, however, lacking the reasoning of contextual relation limits the exploration efficiency. In this work, we propose the self-supervised generative map (SGM), a modular method that learns the explicit context relation via self-supervised learning. The SGM is trained to leverage both episodic observations and general knowledge to reconstruct the masked pixels of a cropped global map. During navigation, the agent maintains an incomplete local semantic map, meanwhile, the unknown regions of the local map are generated by the pre-trained SGM. Based on the generated map, the agent sets the predicted location of the target as the goal and moves towards it. Experiments on Gibson, MP3D and HM3D show the effectiveness of our method. The code is available at https://github.com/sx-zhang/SGM.

1. Introduction

Navigating to specified targets [27, 44, 45, 55] is an essential capability for embodied AI systems to effectively manipulate and interact with real-world entities. Consequently, the visual object goal navigation (ObjectNav) task has recently gained widespread attention. In ObjectNav task, the agent is placed in an unseen and unmapped environment, and is tasked to navigate to an object of the user-specific category (e.g. couch) based on the visual observations. Since the environment is unseen, when the target is invisible, the agent needs to infer the likely location of the target. This requires the agent to learn and understand the contextual relationships between objects (e.g. couches typically appear with cushions and coffee tables), which enables it to deduce the target's location based on observed visual clues.

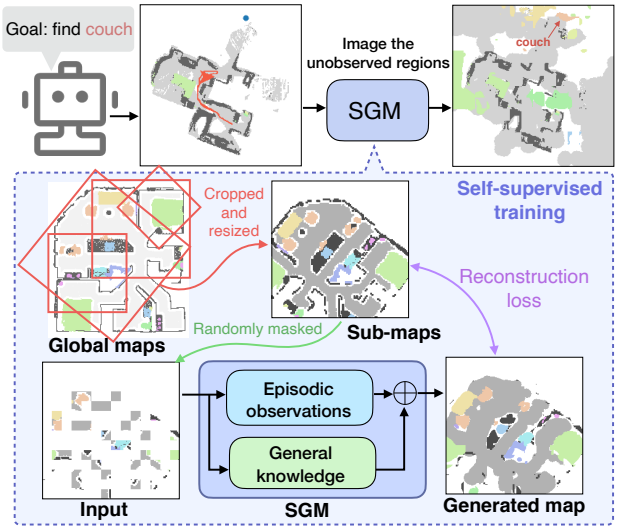

Figure 1. Agents employ SGM to 'imagine' details of unobserved regions during navigation. The SGM is trained in a self-supervised manner by predicting pixels in masked regions of sub-maps to capture contextual relations. The SGM leverages both episodic observations and general knowledge for map reconstruction.

To learn the contextual relations, end-to-end reinforcement learning (RL) methods embed pre-constructed object relation graph [52, 54, 59] into the end-to-end navigation models, or learn object associations directly via the RL [14, 15, 33]. These pre-learned priors are implicitly encoded within the policy. However, since the priors are episode-agnostic and lack geometric memory of current environment, the end-to-end RL methods exhibit limited generalization in unseen environments [32, 60]. Alternatively, modular methods [5–7] construct an explicit semantic local map, merging observed geometry and semantics of the unseen environment. The agent can disregard the observed

regions of the local map, where the target has not appeared, which simplifies the potential state space that needs to be explored. However, without learning the contextual object relation, the modular methods are still limited.

Recently, some methods [37, 56, 61] attempt to introduce priors of object contextual relationships into modular methods for predicting the long-term goal. They learn the contextual relation based on the local map by predicting the potential frontier nearest to the target [37], estimating the distance to the target [61], or directly anticipating the coordinate of the target [56]. However, individually learning the relations of targets with supervised learning may not be reliable, as the absolute position of targets may vary with different room layouts. In contrast, the joint associative relations among multiple objects are more reliable, e.g. chairs, tables, and cups are commonly found together, similarly, couches, cushions, and coffee tables often appear in conjunction. Therefore, we propose learning the joint contextual relations of the objects in a self-supervised manner: without any need for data collection and labeling as supervised learning, the model is trained to reconstruct the masked regions of a cropped global map based on its visible neighboring regions. This preparatory phase allows it to grasp not only individual object specifics (e.g. location and size) but also the patterns of its contextual companions (i.e. contextual object relation) and the surrounding environments (e.g. potential obstacle and free path). Then the trained model can 'imagine' unobserved regions during navigation to help agent deduce the target location.

In this paper, we propose self-supervised generative map (SGM) to scale up the coverage of the local map by generating the unobserved regions, as shown in Fig. 1. Specifically, given the global map of a training room, the global map is cropped into sub-maps with boxes of various scales, positions and angles, then uniformly resized to a fixed scale. Following the self-supervised learning settings of MAE [22], certain patches of the sub-maps are randomly masked. Our SGM is trained to predict the pixel values of the masked patches by leveraging both episodic observations and general knowledge. The episodic observations refer to the visible patches, which are encoded with a visual encoder, akin to many visual pre-training studies [8, 22]. However, contrary to image datasets, the quantity and diversity of training scenes for ObjectNav is notably limited. To enhance the generalization, general knowledge provided by Large Language Models (LLMs) is also considered. Visible patches are transformed into prompts for LLMs through a textual template, subsequently, LLMs (e.g. GPT-4 [34], ChatGLM [17]) predict the semantic contents of the masked patches in text. Furthermore, text predictions (general knowledge) and episodic observations merge via a cross-modality encoder. The decoder predicts the pixel values of the masked patches based on the encoded information. Episodic observations and general knowledge are orthogonal and complementary, where episodic observations provide more geometric details, while general knowledge with rich semantic prior enhances the inference of contextual categories.

During navigation, the agent maintains a local semantic map that integrates observations of current unseen environment. Since the local map is incomplete, the trained SGM is employed to generate the unobserved regions of the local map. The local map is initially divided into non-overlapping patches. Then we propose a sample strategy that prioritizes the patches, which are informative or adjacent to unobserved regions. The selected patches are input into the SGM to predict the pixel values of the unknown regions. The agent selects the coordinates with the highest confidence of the target in the generated map as the long-term goal. The SGM continually generates the unknown regions based on updated local map until the target is found. We evaluate our SGM on photorealistic 3D environments of Gibson [48], Matterport3D (MP3D) [4] and Habitat-Matterport3D (HM3D) [49]. The experimental results demonstrate the effectiveness of our SGM in generating unknown regions and assisting the agent in inferring the location of the target.

2. Related Works

ObjectGoal navigation. Existing visual object goal navigation can be categorized into: end-to-end methods and modular methods. The end-to-end methods learn to navigate by leveraging reinforcement learning (RL) [47, 58, 62] or imitation learning (IL) [38, 39]. Previous end-to-end RL methods attempt to learn object relation graph [14, 52, 54, 59], visual representation [24, 26, 33], historical states representation [16], auxiliary tasks [53] and data augmentation [10, 32] to enhance the navigation ablility. To learn the contextual object relation for enhancing the deducing for target location, [52] constructs prior knowledge of object relations from external datasets [29], and [59] builds hierarchical object-to-zone graph. [14] proposes learning the object relations (i.e. the edges of predefined graph) by RL process. The learned contextual relations are implicitly encoded into the end-to-end model. However, lack of geometry memory and low sample efficiency of RL, the generalization of these methods in unseen environments remains limited. The modular methods [5–7] maintain a geometry semantic map for localization, memory and path planning. Recent modular methods [37, 56, 61] aim to address 'where to look?' subproblem (i.e. inferring the likely location of the target). These methods employ supervised learning to learn a target-related function. Based on a local map, they predict the nearest frontier to target [37], the minimal distance to target [61] or the absolute coordinates [56] of the target. Our SGM is also a modular method, while we learn joint contextual relations, i.e. the SGM predicts the unobserved regions, which contain both objects (not limited to

target) and environments with self-supervised learning.

Several recent works attempt to predict the unobserved regions. [36] proposes to anticipate occupancy for point goal navigation. [19] takes a further step by adding the semantic prediction via two-stage segmentation models. [28] also predicts the unobserved semantic map, while implicitly encoding it into an RL-based policy. All of these works [19, 28, 36] only predict unobserved regions of a top-down map projected from the egocentric RGB-D view at a single timestamp. However, our SGM has a wider prediction scope as it predicts based on a broader local map, which integrates all historical observations. Additionally, it is worth to highlight that our SGM is trained through self-supervised learning, which avoids the tedious data collection and annotation of imitation and supervised learning.

Self-supervised learning. Reconstructing signals from masked inputs has been validated as an effective self-supervised learning within the communities of NLP [2, 12, 35], CV [1, 13, 22], and multimodal pretraining [43]. We adopt the settings of Masked Autoencoders (MAE) [22] to train our SGM, which has demonstrated its robust capacity for learning contextual associations in many visual tasks [8, 18, 51]. Recently, there are few self-supervised works for goal-oriented navigation tasks. [20] proposes a distance estimator from passive videos through self-supervised learning for image goal navigation. Our SGM is proposed for ObjectNav task, which learns the contextual object relation by self-supervised learning, thereby generating the map of unobserved regions during navigation.

3. Approach

3.1. Task Definition

The ObjectNav task is defined as: the agent is required to navigate to an instance of a specified object category (e.g. potted plant) in an unseen environment. The agent is placed at a random location at the start of an episode. At each timestamp t, the agent receives egocentric RGB-D observations s_t, target object o and senor pose (x_t, y_t, θ_t), where x_t, y_t and θ_t denote coordinates and orientation of the agent. The agent executes a discrete action, where the action space consists of `move_forward`, `turn_left`, `turn_right` and `stop`. The agent autonomously executes the action `stop` when it determines to complete the task. A successful episode is denoted as that, within certain number of steps, the agent stops at a position where the distance to the target is less than a threshold (e.g. $1m$) and the target is visible in the egocentric observation.

Our method is constructed based on the modular ObjectNav architecture. The modular methods typically build an accumulating semantic map during navigation [5] based on the RGB-D observations, sensor poses and segmentation models [21, 23]. The semantic map $m_t \in \mathbb{R}^{(N_o+N_s)\times H\times W}$

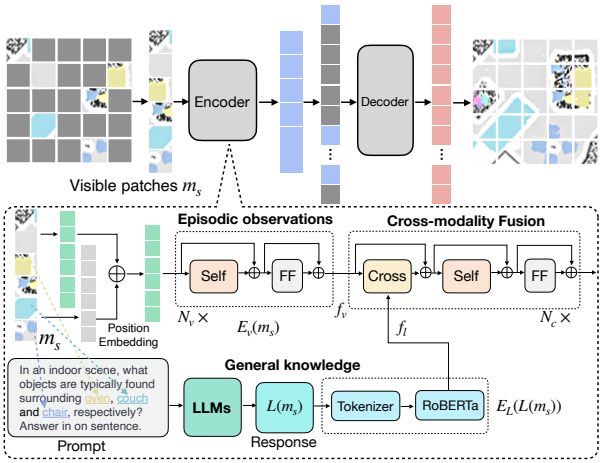

Figure 2. Self-supervised training of SGM. 'Self', 'Cross' and 'FF' denotes self-attention, cross-attention and feed-forward layers, respectively. The SGM is trained to predict the pixel value of the masked patches by utilizing episodic observations and the general knowledge from LLMs. Examples of responses from LLMs are detailed in the supplements.

utilizes multiple channels to represent different classes, where N_o denotes the number occupancy classes including *occupied* and *free*, N_s is the number of object classes and H, W are the map size. The semantic map aggregates the observed object layout of the unseen environment from 0 to t. However, since the semantic map is obtained only by partial observations of the entire environment, it is only a subset (i.e. local map) of the global map. Therefore, when the target is invisible, the agent needs to estimate its unobserved surroundings based on the learned context relation of the objects to infer the likely location of the target.

3.2. Self-supervised Generative Map

To predict the potential position of the target, the contextual relations between objects can be used to generate the unseen regions of local maps. In particular, we propose the self-supervised generative map (SGM) to learn the context relations of both objects and environments (e.g. obstacle and free path) with the self-supervised setting of MAE [22].

As shown in Fig. 1, given a global map of a training room, we first crop and sample a number of sub-maps from it using several boxes in various scales, positions and angles. Then the sub-maps are resized to a fixed scale and divided into non-overlapping patches. These patches are sampled without replacement by a uniform distribution strategy. The selected patches serve as the visible patches m_s, while the remaining patches m_u are masked.

To learn the contextual relation, the SGM $P(m_u|m_s)$ is trained in a self-supervised manner by reconstructing the masked patches m_u based on their visible neighboring patches m_s. As shown in Fig. 2, the proposed SGM reconstructs the masked patches by leveraging the following

two aspects of information: 1) episodic observations, and 2) general knowledge from large language models (LLMs).

Episodic observations. The visible patches are embedded with a linear projection and added with positional embeddings, and then processed with the visual encoder E_v. The visual encoder is implemented by several Transformer blocks of ViT [13], consisting of self-attention and feed-forward layers. Then the episodic observations are embedded as $E_v(m_s) = f_v \in \mathbb{R}^{n_v \times n_d}$, where n_v is the number of visible patches and n_d is the feature dimension. Following [22], the encoder only operates the visible patches, rather than all patches. Therefore, when the ratio of visible patches is small (e.g. 25%), the encoder consumes minimal computing and memory resources.

General knowledge. Due to limited training environments, the quantity and diversity of available global maps for training is considerably less than that of image datasets [11, 29]. As a result, models trained merely on visual clues are prone to overfitting to the limited training environments, leading to poor generalization to unseen environments [32]. Currently, the LLMs demonstrate strong reasoning abilities on many complex tasks [3]. In our work, we employ the LLMs (e.g. GPT-4 [34], ChatGLM [17]) to provide general knowledge for reasoning the contextual relation and predicting the probable objects, thereby enhancing generalization for unseen environments.

For all visible patches, an object category is considered as observed if there is a pixel with the value of 1 existing in the corresponding channel. Then we fill in the observed categories into a fixed sentence template to generate the text prompt for LLMs. For instance, when couches and chairs are observed, the generated prompt is 'In an indoor environment, if couches and chairs are observed, what other objects might be around them? Answer in one sentence.'. The LLMs receive the text prompt and predict the probable neighboring categories around the observed categories. The predictions $L(m_s)$ of LLMs are further processed by Tokenizer and embedded with pre-trained RoBERTa [30] by $E_L(L(m_s)) = f_l \in \mathbb{R}^{n_l \times n_d}$, where n_l is the token number. Since LLMs are trained on massive data, the prediction f_l can be regarded as an episode-agnostic, general knowledge. General knowledge f_l provides a wealth of semantic priors, which is orthogonal and complementary to episodic observation f_v that captures geometric details.

Fusion and anticipation. We utilize the cross-modality encoder of LXMERT [42] to fuse f_v and f_l, which consists of cross-attention, self-attention and feed-forward layers. The cross-attention layers are used to fuse the information between f_v and f_l, which is formally given by

$$CrossAttn\left(f_v, f_l\right) = \sigma\left(\frac{f_v W_q \left(f_l W_k\right)^T}{\sqrt{n_d}}\right) f_l W_v \quad (1)$$

where σ is the Softmax activation function, and $W_* \in \mathbb{R}^{n_d \times n_d}$ are learned parameters and biases are omitted. The output of the cross-modality encoder serves as the final embedding for all visible patches.

The decoder takes full patches (i.e. encoded visible patches and masked patches) as the inputs, where the masked patches are initialized to the learnable vectors of the same size as the encoded visible patches. All patches are added with positional embeddings that represent their spatial locations. The decoder is also implemented by a series of Transformer blocks.

Training. The SGM is trained to predict the pixel value of the masked patches. Given the prediction m_u, the training objective for all classes (i.e. occupancy and object classes) is given by

$$L_{BCE} = \frac{1}{n_u} \sum_{i=0}^{n_u} BCE\left(m_u\left(i\right), \hat{m}_u\left(i\right)\right) \quad (2)$$

where BCE denotes the pixel-wise binary cross entropy loss and n_u is the number of masked patches. $m_u(i) \in \mathbb{R}^{(N_o+N_s)\times\delta\times\delta}$ represents one of the predicted patches, where δ is the size of a patch. $\hat{m}_u\left(i\right)$ is the ground-truth of the masked patches, and each pixel of $\hat{m}_u(i)$ is a binary value. Additionally, to enhance the accuracy of predicting the position and size of objects, an IoU loss is added specifically for the object classes

$$L_{IoU} = \frac{1}{n_u} \sum_{i=0}^{n_u} IoU\left(m_u\left(i, N^s\right), \hat{m}_u\left(i, N^s\right)\right) \quad (3)$$

where IoU denotes the category-level pixel-wise intersection over union loss, and $m_u(i, N^s) \in \mathbb{R}^{N_s \times \delta \times \delta}$ only contains the object classes. We add up all the losses to train our SGM. The overall training objective is $L = L_{BCE} + \lambda L_{IoU}$, where λ denotes the loss weights. The advantages of our SGM are as following: 1) Data collection. The self-supervised training eliminates the requirement for target-oriented annotations or supervision from simulators. SGM requires only training with cropped and randomly masked global maps. 2) Scalability. SGM is compatible with various LLMs, allowing for concurrent improvements in our method as LLMs advance.

3.3. ObjectNav with SGM

The trained SGM is capable of generating unobserved regions based on their visible surroundings, which is utilized to help agent deduce the unobserved regions of the local maps during navigation.

Unobserved regions generation. At each timestamp t during navigation, the agent constructs a local semantic map m_t (as introduced in Sec. 3.1), and each channel of m_t contains binary-valued pixels. The local semantic map includes numerous unobserved regions as shown in Fig. 3, where the pixel values of these regions are 0 across all channels.

Given the local semantic map m_t of time t, we first crop a sub-region of m_t, where the size of this region is ϵ times that of the smallest fitting square box around the known regions of m_t as shown in Fig. 3. The ϵ (e.g. $\epsilon = 140\%$ is decided via validation experiments) represents a scaled-up factor with a minimum value of 100%. The cropped region is resized to a fixed scale $m'_t \in \mathbb{R}^{(N_o+N_s)\times L\times L}$ (e.g. $L = 224$ for ViT-Base) and divided into non-overlapping patches. The size of each patch is δ. To eliminate noise and reduce computational complexity, we select a subset of patches to serve as the input of the SGM. The sampling strategy prioritizes the patches, which are 1) informative and 2) adjacent to unobserved regions.

To construct the sampling strategy, we calculate the average pixel value in each patch by $W = Avgpool(m'_t)$, where $W \in \mathbb{R}^{\frac{L}{\delta}\times\frac{L}{\delta}}$, and $Avgpool$ represents the average pooling in each patch for all channels. For each element $w_i \in W$, the higher value indicates that the corresponding patch is more informative, i.e. there are more observed object classes or larger observed regions in this patch. Furthermore, we leverage the Laplacian kernel K, a gradient operator (i.e. edge detector), to operate the W by $V = Conv(W, K)$, where $V \in \mathbb{R}^{\frac{L}{\delta}\times\frac{L}{\delta}}$ and $Conv$ denotes the convolution operation. The higher value of $v_i \in V$ demonstrates that the corresponding patch is more likely to be adjacent to the unobserved regions.

The W and V are further processed by flattening and Softmax operations for probability normalization. Then the sampling probability P of each patch is defined as

$$P = \alpha\bar{W} + (1-\alpha)\bar{V} \tag{4}$$

where $\bar{W}$ and $\bar{V}$ denote the normalized W and V, respectively, and α is a trade-off weight. We employ a multinomial distribution for sampling patches. Specifically, given a set of patches $m'_t = \{m'_t(1), m'_t(2), \ldots, m'_t(n_p)\}$ ($n_p = (\frac{L}{\delta})^2$), and their associated probabilities $P = \{p_1, p_2, \ldots, p_{n_p}\}$. The patches are sampled without replacement according to $m'_t(i) \sim Multinomial(n, P)$. The number of sampled patches n constitutes only a small fraction (e.g. 30%) of the total.

Based on this strategy, sampled patches are fed into the SGM to predict unobserved regions. The predicted values are continuous, which reflect the confidence in its prediction. Furthermore, the raw prediction of SGM is enhanced by adding the actual local semantic map to form the final generated map, where if a coordinate has already been observed (having values in semantic map), its predicted value is replaced with the confirmed observed value. During ObjectNav task, the agent identifies the pixel with the highest confidence in the target's channel of the generated semantic map. Then the coordinates of this pixel are set as the long-term goal g_t. Additionally, limited observation (e.g. no available contextual object is observed) may result in unreliable prediction of the target (i.e. the confidence of the predicted semantic map is low). Therefore, we adopt the exploration strategy [31, 50] at the beginning of navigation until the prediction confidence of the target exceeds a certain threshold.

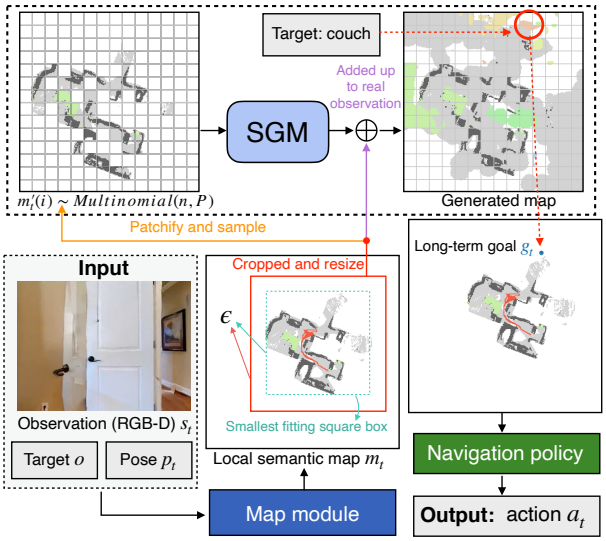

Figure 3. During navigation, the agent builds a local semantic map m_t and crops it to m'_t. The patches of m'_t are sampled by multinomial distribution with probability P. The SGM generates the unobserved regions based on selected n patches. The long-term goal is determined by the generated map and the target.

Navigation policy. Once the long-term goal is identified, the agent simply needs to navigate from its current location (x_t, y_t) to the long-term goal g_t. Following previous works [5, 37], the local policy is implemented by the Fast Marching Method [41], which calculates the shortest path from current location to the long-term goal based on the occupancy channels of the generated map. Then the local policy calculates deterministic actions for the agent based on its step distance and the shortest path. At each timestamp, the local policy will re-plan the actions for the agent according to the updated semantic map.

4. Experiments

4.1. Experimental Setup

Dataset. We evaluate our SGM on standard ObjectNav datasets, including Gibson [48], Matterport3D (MP3D) [4] and Habitat-Matterport3D (HM3D) [49] dataset.

For Gibson and MP3D, we follow the setup of [37, 57], we we utilize 25 train /5 val scenes from the Gibson tiny split and choose 6 goal categories with 1,000 val episodes for Gibson. For MP3D, we employ 56 train / 11 val scenes, 21 goal categories and 2,195 episodes for validation. For HM3D, our setup is consistent with [9], where we choose 80 train / 20 val scenes, 6 goal categories and 2,000 valida-

Table 1. Comparisons of different map generation variants in the Gibson (val). R/C means the training map is masked by random (R) or connected (C) strategy. The connected masking strategy requires that retained patches must be adjacent to each other (remaining patches are akin to a local map), which is similar to [19, 28] that learn contextual relations by completing the local map. For LLMs, we compare GPT-4 [34] and ChatGLM [17].

ID		Model	Mask	LLM	Map Generation		Navigation		
					IoU(%)	Recall(%)	SR(%)	SPL(%)	DTS(m)
I	1	UNet	-	-	19.28	47.36	67.5	37.8	1.76
II	2	ViT-L	C	-	24.18	59.42	73.8	39.5	1.45
	3	ViT-L	R	-	**26.85**	79.05	**74.2**	**39.6**	1.49
	4	ViT-B	C	-	22.31	55.50	73.1	39.3	1.53
	5	ViT-B	R	-	25.45	**79.23**	74.1	39.5	**1.48**
III	6	ViT-B	C	GPT-4	32.06	72.34	76.8	43.4	1.27
	7	ViT-B	R	GPT-4	**39.47**	87.33	77.4	43.8	1.15
	8	ViT-B	C	ChatGLM	30.87	80.67	77.0	43.5	1.31
	9	ViT-B	R	ChatGLM	31.73	**91.68**	**78.0**	**44.0**	**1.11**

tion episodes. The goal categories adopted by these three datasets are listed in supplements.

Evaluation metrics. For evaluating the navigation performance, we adopt three standard metrics following [5, 37, 56]: 1) **SR**: the ratio of success episodes. 2) **SPL**: the success rate weighted by the path length, which measures the efficiency of the path length. 3) **DTS**: the distance to the goal at the end of the episode.

To assess the quality of map predictions, we employ the following two metrics: 1) **IoU**: the pixel intersection over union for all channels (both occupancy and object). 2) **Recall**: the proportion of true positive predicted object categories over the total number of ground-truth categories.

Implementation details. For the self-supervised training of SGM, we sample 400,000 train /1,000 val sub-maps for training and validation in each dataset. We implement the SGM based on ViT-base/16, where the visual encoder, cross-modality encoder and decoder are respectively composed of 8, 4 and 2 Transformer blocks. The training mask ratio is set to 75%, the input size $L = 224$, and patch size $\delta = 16$. We use the Adam optimizer [25] with a base learning rate of 0.00015. The warmup epoch is linearly scaled when the training epoch increases to 40. After warmup epochs, the learning rate is decayed by a factor of 0.05. The loss weights $\lambda = 0.2$.

For ObjectNav task, since constructing local semantic map requires the semantic segmentation model, we adopt publicly available Mask-RCNN [21] from [37] for Gibson, RedNet [23] from [37] for MP3D, and Mask-RCNN [21] from [56] for HM3D. The scaled-up factor $\epsilon = 140\%$, the number of selected patches $n = 59$, and sample weights $\alpha = 0.5$. The experiments for these hyper-parameters are detailed in supplements. The turn angle is set to 30 degrees, and the step distance of `move_forward` is 25cm.

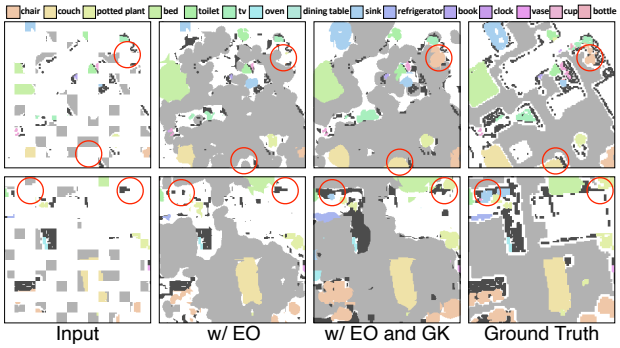

Figure 4. Map reconstruction results. The 'w/ EO' means only using episodic observations in reconstruction, while 'w/ EO and GK' denotes leveraging both episodic observations and LLMs (ChatGLM). General knowledge enables the model to predict the completely unobserved contextual objects (see the regions highlighted by red circles). These sub-maps sourced from the val rooms of Gibson are unseen during training.

4.2. Evaluation Results

Comparisons with previous map completion methods. Prior methods [19, 28] train a UNet [40] to directly predict unobserved regions based on the top-down map of a single timestamp. This comparison is to determine whether using UNet or learning contextual relations directly from the local map is more effective than our SGM (utilizing ViT and learning with randomly masked sub-maps). As shown in Tab. 1, we compare these two variants under identical training epochs and datasets. The UNet variant underperforms others in both map generation and navigation. In learning contextual relations, the random masking strategy (SGM) outperforms connected masking strategy [19, 28]. We infer that since visible patches are closely clustered in the connected masking strategy, the context of patches distant from the visible ones is completely absent, leading to difficulties for model in capturing the context relations of these patches during training. The results demonstrate that our SGM is more effective in capturing contextual relationships and assisting the navigation.

Model complexity. We compare the impact of more complex models on map generation and navigation. As indicated by the row II in Tab. 1, the more complex model (i.e. ViT Large) only yields a marginal performance improvement. Hence, we conjecture that the ViT Base model is sufficiently capable for the task. Consequently, our SGM is ultimately implemented based on the ViT Base model.

Impact of LLMs. Our SGM not only leverages episodic observations but also incorporates general knowledge obtained from LLMs for prediction. Compared to rows II and III in Tab. 1, there is a significant improvement on the recall metric, which suggests that LLMs can substantially bolster the prediction of object categories. As shown

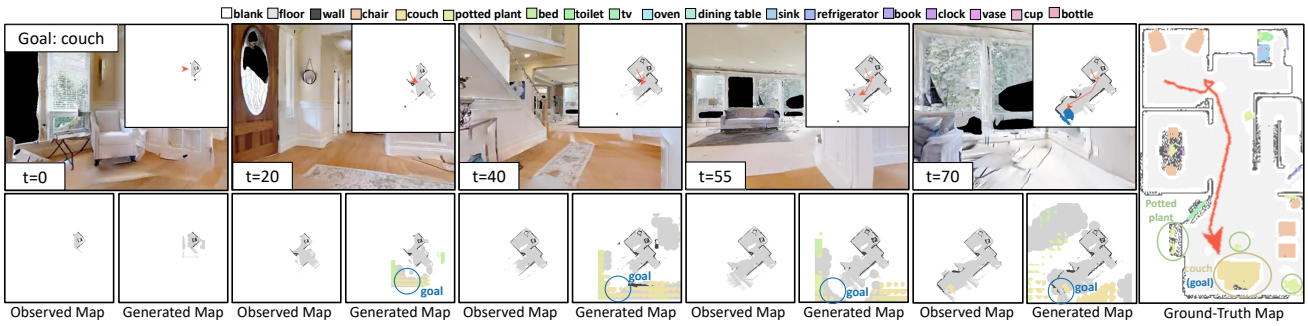

Figure 5. Navigation episode with SGM in Gibson (val). The top row displays the agent's RGB view and local semantic map, including the trajectory and long-term goal (indicated by the deep blue dot). The bottom row shows the generated map by SGM. At about 30% of the navigation process (i.e. t=20), the generated map accurately predicts the target's location even though the target has not yet been observed. Besides, SGM not only precisely predicts the target (couch), but also its contextual objects (e.g. potted plant).

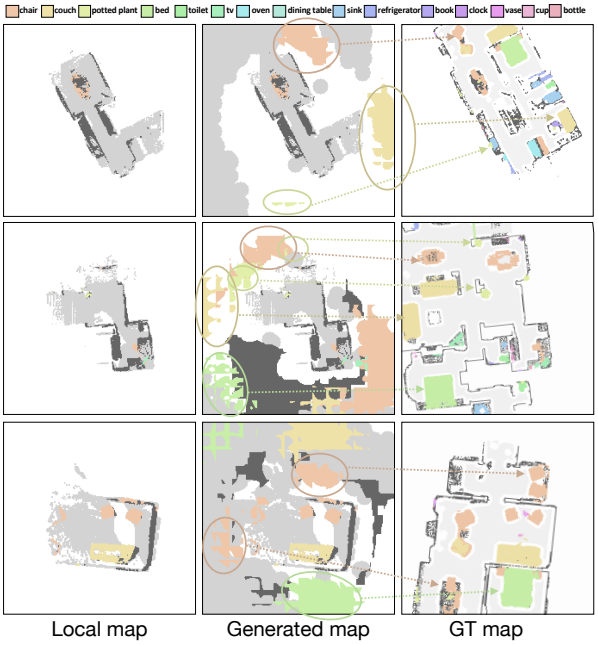

Figure 6. More generated maps during navigation. As shown by the circled objects, the generated map accurately anticipates the location of unobserved objects.

in the Fig. 4, the model with solely episodic observations is capable of completing the contours of partially observed objects, while it struggles to predict entirely unobserved categories. However, the general knowledge from LLMs provides rich semantic contextual priors, enables the model to predict the context objects that are completely unobserved (e.g. the chairs in the first row, and the potted plants and sinks in the second row). Additionally, we compare two LLMs: GPT-4 [34] and ChatGLM [17]. The results show that the choice of LLMs has minimal impact on performance. Moreover, as ChatGLM is publicly available, thus, we employ it to provide general knowledge. To improve inference speed, we extract all possible combinations of goal categories to pre-generate the prompt and obtain responses by using LLMs. The SGM directly leverages pre-extracted responses for prediction during both training and testing.

Table 2. Ablation study of different components. GT means the long-term goal is set to the ground truth location of the target. 'EO' denotes the episodic observation, 'GK' represents the general knowledge and 'Explr' means using exploration strategy at the beginning of navigation.

ID		Modules				Gibson			MP3D		
		GT	EO	GK	Explr	SR(%)	SPL(%)	DTS(m)	SR(%)	SPL(%)	DTS(m)
I	1	✓				91.7	72.3	0.38	-	-	-
	2		✓			71.5	40.8	1.46	31.9	12.2	5.32
	3		✓	✓		74.4	41.3	1.49	34.7	13.2	5.60
II	4				✓	65.1	37.9	1.76	25.9	11.6	5.62
	5		✓		✓	74.1	39.5	1.48	33.0	11.8	5.10
	6		✓	✓	✓	**78.0**	**44.0**	**1.11**	**37.6**	**14.7**	**4.93**

Visualization of navigating with SGM. We further visualize the generated maps of SGM during the navigation, as shown in Fig. 5 and 6, noting that the scenes are unseen. As Fig. 5 illustrated, at the beginning of navigation ($t = 0$), the generated image offers limited additional information due to minimal observations. However, as more of the scene becomes observed, by $t = 20$ (only about 30% of the entire navigation), SGM successfully predicts the location of the target (couch) even though it has not yet been observed. Then, SGM continues to guide the agent by determining the long-term goal based on the generated map. Moreover, it is notable that SGM not only accurately locates the target but also anticipates its context objects (e.g. potted plant). Fig. 6 provides additional generated results, which is evident that the generated maps correctly estimate the approximate locations of objects and their context, while there is still a deviation in their absolute positions. However, we consider this deviation to be tolerable, as even humans cannot precisely locate unseen objects in unseen scenes. These two visualizations indicate that our SGM effectively scales up limited local maps by generating the unobserved regions, thereby guiding the agent to infer the target's location.

Ablations study. As shown in row I of Tab. 2, the agent

Table 3. Comparisons with the related works in Gibson and MP3D val. Note that THDA [32] and Habitat-Web [38] use additional data for training, while Red-Rabbit [53] utilizes auxiliary tasks to train the agent. For SemExp [5], L2M [19] and Stubborn [31], we report results from [57]. * denotes our implementation.

	ID	Method	Gibson			MP3D		
			SR(%)	SPL(%)	DTS(m)	SR(%)	SPL(%)	DTS(m)
	1	Random	0.4	0.4	3.89	0.5	0.5	8.05
	2	DD-PPO [46]	15.0	10.7	3.24	8.0	1.8	6.94
	3	Red-Rabbit [53]	-	-	-	34.6	7.9	-
	4	THDA [32]	-	-	-	28.4	11.0	5.62
I	5	SSCNav* [28]	-	-	-	27.1	11.2	5.71
	6	EmbCLIP* [24]	68.1	39.5	1.15	29.2	10.1	5.42
	7	Habitat-Web [38]	-	-	-	35.4	10.2	-
	8	ENTL [26]	-	-	-	17.0	5.0	-
	9	FBE [50]	48.5	28.9	2.56	29.5	10.6	5.00
	10	ANS [7]	67.1	34.9	1.66	21.2	9.4	6.31
	11	SemExp [5]	71.1	39.6	1.39	28.3	10.9	6.06
II	12	PONI [37]	73.6	41.0	1.25	27.8	12.0	5.63
	13	L2M [19]	-	-	-	32.1	11.0	5.12
	14	Stubborn [31]	-	-	-	31.2	13.5	5.01
	15	3D-aware [57]	74.5	42.1	1.16	34.0	14.6	**4.78**
	16	SGM (Ours)	**78.0**	**44.0**	**1.11**	**37.7**	**14.7**	4.93

only relying on SGM has already achieved comparable performance to some supervised methods [5, 37]. Furthermore, the row II of Tab. 2 indicates that introducing an exploration strategy in the early stages of navigation is effective, which prevents misguidance by low-confidence predictions resulted from limited observations. Our exploration strategy utilizes the area potential function from [37], which calculates the nearest frontier as the long-term goal. In contrast to the computation-based FBE [50], the area potential function learns to predict potential frontiers through supervised training. The ablation studies demonstrate the effectiveness of each component in our method.

Comparisons with the related works. Since there is no existing self-supervised work for ObjectNav task, our comparison is limited to previous supervised methods. We consider the following baselines: the end-to-end RL methods [9, 24, 26, 28, 32, 46, 53], the imitation learning method [38], and modular methods [5, 6, 19, 31, 37, 50, 56, 57] Note that, some works leverage additional data [10, 32, 38] or auxiliary task [9, 53] to improve the performance. Thus, it is challenging to compare all the methods in an equitable manner. Consequently, our focus is specifically on the most relevant baselines: SemExp [5], PONI [37], L2M [19], SSCNav [28], PEANUT [56], where PONI and PEANUT integrate goal-related function trained by supervised learning into SemExp, while L2M and SSCNav enhance navigation performance by learning to complete egocentric top-down map of single timestamp. They are also trained with supervised learning, while are limited to consider only single-frame observations. In contrast, our SGM predicts the unobserved regions for the broader local map, which incorporates full observations from time 0 to t. Notably, PEANUT [56] initially reports results on MP3D and HM3D, however, they only release their code for HM3D. Therefore, we only compare with [56] on HM3D with re-implementation results in our setup. Besides, L2M and SSCNav are evaluated with their self-made validation data. Thus, for comparing their results, we report our own implementations or adopt re-implementation results from other work [57] with experimental settings that align with ours.

Table 4. Comparisons with the related works in HM3D val. * denotes our implementation.

ID	Method	HM3D	
		SR(%)	SPL(%)
1	DD-PPO [46]	27.9	14.2
2	Habitat-Web [38]	57.6	23.8
3	RIM [9]	57.8	27.2
4	PEANUT* [56]	59.1	30.3
5	SGM (Ours)	**60.2**	**30.8**

We compare our SGM with the related works on the validation set of Gibson, MP3D and HM3D datasets, as shown in Tab. 3 and 4. Despite our SGM being self-supervised, it achieves comparable performance with existing supervised methods across all metrics on these datasets. Particularly on the Gibson dataset, compared to the current state-of-the-art [57], our SGM outperforms 3D-aware [57] by 3.5%, 1.9%, and -0.05m in SR, SPL and DTS metrics, respectively. Note that lower value on DTS indicates better performance.

5. Conclusions

In this work, we present the Self-supervised Generative Map (SGM) for the object goal navigation task. Our SGM captures contextual relationships of both objects and environments through self-supervised training. SGM utilizes two distinct yet complementary information for prediction: episodic observations and general knowledge from Large Language Models (LLMs), where episodic observations provide geometric details, and general knowledge offers semantic priors. During navigation, the local semantic map maintained by agent is sampled by proposed sampling strategy to select informative patches. Then the selected patches are fed into SGM to generate the unobserved regions. The generated map helps agent infer the likely location of the target. The experimental results indicate that despite being a self-supervised method, our SGM achieves comparable performance to previous supervised methods.

Acknowledgements: This work was supported by the National Natural Science Foundation of China under Grant 62125207, 62272443, 62032022 and U23B2012, in part by Beijing Natural Science Foundation under Grant JQ22012, Z190020.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 4

[4] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pages 667–676. IEEE Computer Society, 2017. 2, 5

[5] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Russ R. Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 2, 3, 5, 6, 8

[6] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural SLAM. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 8

[7] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological SLAM for visual navigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12872–12881, 2020. 1, 2, 8

[8] Anthony Chen, Kevin Zhang, Renrui Zhang, Zihan Wang, Yuheng Lu, Yandong Guo, and Shanghang Zhang. Pimae: Point cloud and image interactive masked autoencoders for 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5291–5301. IEEE, 2023. 2, 3

[9] Shizhe Chen, Thomas Chabal, Ivan Laptev, and Cordelia Schmid. Object goal navigation with recursive implicit maps. *CoRR*, abs/2308.05602, 2023. 5, 8

[10] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *NeurIPS*, 2022. Outstanding Paper Award. 2, 8

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009. 4

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3, 4

[14] Heming Du, Xin Yu, and Liang Zheng. Learning object relation graph and tentative policy for visual navigation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII*, pages 19–34, 2020. 1, 2

[15] Heming Du, Xin Yu, and Liang Zheng. Vtnet: Visual transformer network for object goal navigation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 1

[16] Heming Du, Lincheng Li, Zi Huang, and Xin Yu. Object-goal visual navigation via effective exploration of relations among historical navigation states. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2563–2573. IEEE, 2023. 2

[17] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022. 2, 4, 6, 7

[18] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022. 3

[19] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to

map for active semantic goal navigation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3, 6, 8

[20] Meera Hahn, Devendra Singh Chaplot, Shubham Tulsiani, Mustafa Mukadam, James M. Rehg, and Abhinav Gupta. No rl, no simulation: Learning to navigate without navigating. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 26661–26673, 2021. 3

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988, 2017. 3, 6

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. 2, 3, 4

[23] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*, 2018. 3, 6

[24] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: CLIP embeddings for embodied AI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14809–14818. IEEE, 2022. 2, 8

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[26] Klemen Kotar, Aaron Walsman, and Roozbeh Mottaghi. Entl: Embodied navigation trajectory learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10863–10872, 2023. 2, 8

[27] Xiangyang Li, Zihan Wang, Jiahao Yang, Yaowei Wang, and Shuqiang Jiang. Kerm: Knowledge enhanced reasoning for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2583–2592, 2023. 1

[28] Yiqing Liang, Boyuan Chen, and Shuran Song. Sscnav: Confidence-aware semantic scene completion for visual semantic navigation. In *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*, pages 13194–13200. IEEE, 2021. 3, 6, 8

[29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer, 2014. 2, 4

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 4

[31] Haokuan Luo, Albert Yue, Zhang-Wei Hong, and Pulkit Agrawal. Stubborn: A strong baseline for indoor object navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2022, Kyoto, Japan, October 23-27, 2022*, pages 3287–3293. IEEE, 2022. 5, 8

[32] Oleksandr Maksymets, Vincent Cartillier, Aaron Gokaslan, Erik Wijmans, Wojciech Galuba, Stefan Lee, and Dhruv Batra. THDA: treasure hunt data augmentation for semantic navigation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15354–15363, 2021. 1, 2, 4, 8

[33] Bar Mayo, Tamir Hazan, and Ayellet Tal. Visual navigation with spatial attention. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16898–16907, 2021. 1, 2

[34] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 2, 4, 6, 7

[35] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[36] Santhosh K. Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, pages 400–418. Springer, 2020. 3

[37] Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. PONI: potential functions for objectgoal navigation with interaction-free learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18868–18878. IEEE, 2022. 2, 5, 6, 8

[38] Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5163–5173. IEEE, 2022. 2, 8

[39] Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. Pirlnav: Pretraining with imitation and RL finetuning for OBJECTNAV. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 17896–17906. IEEE, 2023. 2

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, pages 234–241. Springer, 2015. 6

[41] James A. Sethian. Fast marching methods. *SIAM Rev.*, 41 (2):199–235, 1999. 5

[42] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 5099–5110. Association for Computational Linguistics, 2019. 4

[43] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19175–19186. IEEE, 2023. 3

[44] Xiaohan Wang, Yuehu Liu, Xinhang Song, Beibei Wang, and Shuqiang Jiang. Camp: Causal multi-policy planning for interactive navigation in multi-room scenes. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[45] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15625–15636, 2023. 1

[46] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: learning near-perfect pointgoal navigators from 2.5 billion frames. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 8

[47] Mitchell Wortsman, Kiana Ehsani, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6750–6759, 2019. 2

[48] Fei Xia, Amir Roshan Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9068–9079. IEEE Computer Society, 2018. 2, 5

[49] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Théophile Gervet, John M. Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-matterport 3d semantics dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 4927–4936. IEEE, 2023. 2, 5

[50] 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, July 10-11, 1997, Monterey, California, USA*, pages 146–151. IEEE Computer Society, 1997. 5, 8

[51] Qingsen Yan, Song Zhang, Weiye Chen, Hao Tang, Yu Zhu, Jinqiu Sun, Luc Van Gool, and Yanning Zhang. Smae: Few-shot learning for hdr deghosting with saturation-aware masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5775–5784, 2023. 3

[52] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 1, 2

[53] Joel Ye, Dhruv Batra, Abhishek Das, and Erik Wijmans. Auxiliary tasks and exploration enable objectgoal navigation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 16097–16106. IEEE, 2021. 2, 8

[54] Xin Ye and Yezhou Yang. Hierarchical and partially observable goal-driven policy learning with goals relational graph. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14101–14110, 2021. 1, 2

[55] Haitao Zeng, Xinhang Song, and Shuqiang Jiang. Multi-object navigation using potential target position policy function. *IEEE Trans. Image Process.*, 32:2608–2619, 2023. 1

[56] Albert J. Zhai and Shenlong Wang. Peanut: Predicting and navigating to unseen targets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10926–10935, 2023. 2, 6, 8

[57] Jiazhao Zhang, Liu Dai, Fanpeng Meng, Qingnan Fan, Xuelin Chen, Kai Xu, and He Wang. 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, 2023. 5, 8

[58] Sixian Zhang, Weijie Li, Xinhang Song, Yubing Bai, and Shuqiang Jiang. Generative meta-adversarial network for unseen object navigation. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIX*, pages 301–320. 2

[59] Sixian Zhang, Xinhang Song, Yubing Bai, Weijie Li, Yakui Chu, and Shuqiang Jiang. Hierarchical object-to-zone graph for object navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15130–15140, 2021. 1, 2

[60] Sixian Zhang, Xinhang Song, Weijie Li, Yubing Bai, Xinyao Yu, and Shuqiang Jiang. Layout-based causal inference for object navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10792–10802. IEEE, 2023. 1

[61] Minzhao Zhu, Binglei Zhao, and Tao Kong. Navigating to objects in unseen environments by distance prediction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2022, Kyoto, Japan, October 23-27, 2022*, pages 10571–10578. IEEE, 2022. 2

[62] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J. Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE International Conference on*

Robotics and Automation, ICRA 2017, Singapore, Singapore, May 29 - June 3, 2017, pages 3357–3364, 2017. 2