# *SAME*: Learning Generic Language-Guided Visual Navigation with State-Adaptive Mixture of Experts

Gengze Zhou[1] Yicong Hong[2] Zun Wang[3] Chongyang Zhao[4] Mohit Bansal[3] Qi Wu[1]
[1]The University of Adelaide [2]Adobe Research [3]UNC, Chapel Hill [4]UNSW Sydney
{gengze.zhou, qi.wu01}@adelaide.edu.au
https://github.com/GengzeZhou/SAME

## Abstract

*The academic field of learning instruction-guided visual navigation can be generally categorized into high-level category-specific search and low-level language-guided navigation, depending on the granularity of language instruction, in which the former emphasizes the exploration process, while the latter concentrates on following detailed textual commands. Despite the differing focuses of these tasks, the underlying requirements of interpreting instructions, comprehending the surroundings, and inferring action decisions remain consistent. This paper consolidates diverse navigation tasks into a unified and generic framework – we investigate the core difficulties of sharing general knowledge and exploiting task-specific capabilities in learning navigation and propose a novel State-Adaptive Mixture of Experts (SAME) model that effectively enables an agent to infer decisions based on different-granularity language and dynamic observations. Powered by SAME, we present a versatile agent capable of addressing seven navigation tasks simultaneously, achieving highly comparable performance to task-specific agents.*

## 1. Introduction

Acquiring the capability to understand natural language commands and navigate in unfamiliar environments constitutes a fundamental competency for embodied intelligence. In recent years, a great variety of navigational tasks has emerged, each defined by distinct navigation objectives, from broad, high-level goals [8, 130] to detailed, low-level directives [6, 54, 81, 101, 129], highlighting exploration and instruction-following, respectively. However, these tasks are mostly formulated as isolated research problems, and the specialized methods developed for each are typically not generalizable to others (Figure 1a). For example, structured memory tailored for efficient target object exploration [10, 11, 88], contextual guidance for vague instructions [33, 66, 97], and episodic vision-language alignment for instruction following [31, 42, 100, 105]. Subsequent works leverage generic vision-language representations [21, 61, 63, 98, 99] to pretrain vision-language-action policies [17, 19, 35, 37, 40, 62, 75, 82] (Figure 1b), finetuning parameters for specific tasks while maintaining the same model architecture. In this paper, we argue that the essential difference between these tasks lies in the granularity of instruction, and a single model's capability should not be constrained by the user's instruction format from the practical perspective. We propose unifying these learning problems under the broader concept of language-guided visual navigation, extending the multi-task learning problem in Vision-and-Language Navigation (VLN) [125] to encompass OBJECTNAV [8] in visual semantic navigation. Our overarching goal is to create a versatile system that can interpret and execute arbitrary language instructions (Figure 1c).

An intuitive approach towards this is to integrate distinct navigational datasets and use them to train an existing agent for multitasking; however, our study reveals that simply mixing data yields inconsistent performance variation across tasks due to conflicting learning objectives (§2.2). Previous work has also identified this challenge and explored generic zero-shot LLM systems without training [71], but such approaches achieved limited results on VLN [12, 13, 15, 16, 64, 70, 84, 110, 126]. Therefore, we aim to build an effective learning paradigm that facilitates training an agent on a larger amount of data covering wider skill sets. As a result, we resort to a more adaptive framework capable of sharing general knowledge and retrieving task-specific skills to infer appropriate navigation decisions.

Inspired by recent success in the Mixture of Experts (MoE) [45, 47, 95] approach for natural language processing, in particular, enhancing the transformer-based Large Language Models [24, 46, 114], we incorporate MoE to build our generalizable navigation agent. Unlike previous works that typically implement task-wise or token-wise MoE, which we found ineffective in multi-task navigation learning, we propose a novel MoE formulation for sequen-

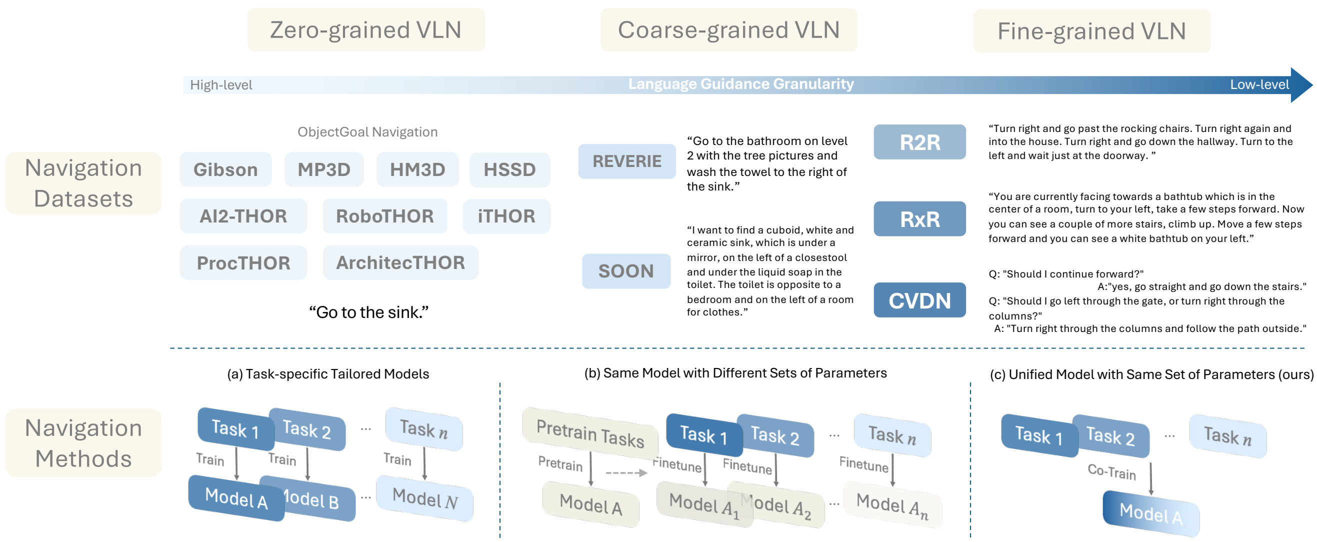


Figure 1. We consolidate diverse navigation tasks into a unified language-guided navigation framework sorted by language granularity. Previous approaches utilize task-specific designs tailored to address particular types of language instructions, as shown in (a) and (b). In contrast, we propose a versatile system that can interpret and execute arbitrary language instructions as shown in (c).

tial embodied agents in which the experts are selected based on the agent's state (i.e., attended language and visual observation at a certain timestep, see §3.3). We further show that applying MoE on visual queries gives better results than applying it on feed-forward networks as in LLMs, aligning with the fact that many navigational agents rely on multi-view visual attention for decision-making.

We termed this method *State-Adaptive Mixture of Experts*, or *SAME*, suggesting applying the *same* model to solve a wide range of navigation problems. Powered by SAME, we are the first to train a versatile agent across seven major language-guided navigation tasks, including R2R [6], RxR-EN [55], REVERIE [81], OBJECTNAV[1], CVDN [101], SOON [129], and R2R-CE [53], achieving state-of-the-art performance as a unified model while matching the performance of task-specific models.

## 2. Background

In this section, we begin by introducing the formulation of language-guided visual navigation tasks under varying levels of language granularity. We employ DUET [19] to illustrate a general cross-modality navigation model for language-guided navigation. Through a series of contrastive experiments, we analyze the interconnections among these navigation tasks, understanding their underlying contradictions and providing a proof of concept for our approach.

[1]We consider Object-Goal Navigation [8] as a form of zero-grained language-guided navigation, as will be specified in §2.1.

### 2.1. Navigation Tasks Formulation

Given an instruction represented by a sequence of $L$ word embeddings $\mathcal{W} = \{w_i\}_{i=1}^{L}$, an agent navigates on a predefined undirected graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ represents the navigable nodes and $\mathcal{E}$ represents the connectivity edges. The agent is expected to execute a sequence of actions $\{s_0, a_0, s_1, a_1, \ldots, s_T, a_T\}$ to navigate to the target position $v_T$, as specified by the instruction $\mathcal{W}$. Each action $a_t$ transit the agent from the current state $s_t = \langle v_t, \theta_t, \phi_t \rangle$ to $s_{t+1} = \langle v_{t+1}, \theta_{t+1}, \phi_{t+1} \rangle$ which includes its spatial location $v_t \in \mathcal{V}$, heading angle $\theta_t$, and elevation angle $\phi_t$, and generate a new visual observation $\mathcal{O}_t$. Additionally, the agent maintains a record of the state history $h_t$ and adjusts the conditional transition probability between states, defined as $\mathcal{S}_t = \mathcal{T}(s_{t+1} \mid a_t, s_t, h_t, \mathcal{O}, \mathcal{W})$, where $\mathcal{T}$ denotes the conditional transition probability distribution. We categorize the language instruction $\mathcal{W}$ into three classes by granularity as:

- **Fine-grained VLN**: $\mathcal{W}$ describes the sequence of actions $\{s_0, a_0, s_1, a_1, \ldots, s_T, a_T\}$ step-by-step.
- **Coarse-grained VLN**: $\mathcal{W}$ refers to a remote target at $v_T$, *e.g.*, "the cold tap in the first bedroom on level two".
- **Zero-grained VLN**: $\mathcal{W}$ refers to a single term indicating the target (*e.g.*, an object category in OBJECTNAV).

**Multimodal Navigation Policy** At each step, the agent receives a local visual observation $\mathcal{O}_t = \{o_i\}_{i=0}^{36}$, consisting of 36 view images, and a language instruction $\mathcal{W}$. These are encoded separately by a vision encoder and a language encoder into visual feature $\hat{\mathcal{O}}_t$ and language feature $\hat{\mathcal{W}}$. DUET [19] incorporates $\hat{\mathcal{O}}_t$ and agent's state $s_t$ to obtain node embedding $\hat{V}_t$ and maintain a topological map

$\hat{\mathcal{G}}_t = \{\hat{V}_i\}_{i \leq t}$ as navigation history, details are provided in the supplementary. A local cross-modal encoder is utilized to excite visual features conditioned on language features:

$$\text{CrossAttn}(\hat{\mathcal{O}}_t, \hat{\mathcal{W}}) = \text{Softmax}\left(\frac{\hat{\mathcal{O}}_t W_q (\hat{\mathcal{W}} W_k)^T}{\sqrt{d}}\right) \hat{\mathcal{W}} W_v. \tag{1}$$

The output embedding from the final layer of view $o_i$ is denoted as $\hat{o}_i'$. Similarly, a parallel global cross-modal encoder is implemented to encode language-conditioned map $\hat{\mathcal{G}}_t{}' = \text{CrossAttn}(\hat{\mathcal{G}}_t, \hat{\mathcal{W}})$. Denote the output embedding of node $V_i$ as $\hat{v}_i'$, the navigation score given by the local and global cross-modal encoder is calculated as:

$$s_i^l = \text{FFN}^l(\hat{o}_i'), s_i^g = \text{FFN}^g(\hat{v}_i'), \tag{2}$$

$$s_i = \sigma_t s_i^l + (1 - \sigma_t) s_i^g, \tag{3}$$

where FFN is a two-layer feed-forward network and $\sigma$ is a learnable parameter.

**Datasets** We select three typical datasets for our contrastive experiments based on instruction granularity:

- R2R [6]: The fine-grained VLN task which consists of 22k human-annotated navigational instructions. On average, an instruction contains 32 words, and the ground-truth path is formed by 7 steps, totaling 10 meters.
- REVERIE [81]: The coarse-grained VLN task inherits the trajectories from R2R but provides high-level instructions that describe a target object. On average, instructions contain 21 words, and the length of ground-truth paths ranges from 4 to 7 steps.
- OBJECTNAV-MP3D [8]: We use the standard split of 11 validation scenes from the Habitat OBJECTNAV dataset [94] in MP3D [9], which consists of 21 goal categories. We utilized human demonstration from Habitat-Web [89] as training data, discussed in Section 2.2.

**Evaluation Metrics** We follow 5 standard metrics in VLN literature to assess the agent performance, including Trajectory Length (TL), Navigation Error (NE), Success Rate (SR), Success Rate normalized by Path Length (SPL) [5], normalized Dynamic Time Warping (nDTW) [44].

## 2.2. What are the Conflicts in Navigation Multi-tasks Learning?

To successfully train a versatile navigation agent, it is essential to understand the underlying contradictions that prevent unified model learning and to gain insights into the best practices for learning from diverse data sources. Following the discussion in Section 1, we unify the training data by transferring 70k human demonstration OBJECTNAV data from Habitat-Web in the discrete environment.

**Data Transformation in Discrete Environment** The original trajectories in OBJECTNAV are constructed as sequences of continuous viewpoint positions, averaging 243 steps per demonstration. Following Hong *et al.* [41], we discretize the Habitat-MP3D [93] environment into a connectivity graph $\mathcal{G}^*$. We match each viewpoint in the trajectory from Habitat-Web to the nearest nodes on $\mathcal{G}^*$ based on Euclidean distance and merge sequentially repeated nodes. Disconnected paths and paths with an ending position more than $0.5m$ away from the original endpoint are removed. This results in 58,803 trajectories with an average of 20 steps. Similarly, we transfer the data from the MP3D validation split to evaluate model performance in discrete environments.

| Training Data | | | R2R (Val Unseen) | | | | REVERIE (Val Unseen) | | | | OBJECTNAV-MP3D (Val) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| R2R* | REVERIE* | MP3D* | TL | NE↓ | SR↑ | SPL↑ | TL | NE↓ | SR↑ | SPL↑ | TL | NE↓ | SR↑ | SPL↑ |
| ✓ | | | 14.33 | 3.82 | 67 | 55 | 19.61 | 7.55 | 39 | 28 | 15.30 | 4.69 | 55 | 24 |
| | ✓ | | 17.55 | 6.22 | 42 | 32 | 17.91 | 6.56 | 41 | 32 | 10.46 | 5.91 | 43 | 23 |
| | | ✓ | 20.76 | 8.55 | 16 | 9 | 20.00 | 10.11 | 13 | 9 | 22.17 | 3.67 | 68 | 29 |
| ✓ | | ✓ | 14.03 | 4.01 | 64 | 55 | 15.22 | 7.78 | 38 | 31 | 25.91 | 3.28 | 72 | 28 |
| | ✓ | ✓ | 19.17 | 7.13 | 34 | 26 | 19.46 | 6.24 | 35 | 26 | 21.50 | 3.29 | 70 | 33 |
| ✓ | ✓ | ✓ | 14.21 | 4.10 | 65 | 54 | 16.62 | 6.11 | 34 | 27 | 22.97 | 3.54 | 68 | 27 |

Table 1. Comparison of single-run performance with a different mixture of training data for DuET [19]. $*$ indicates utilizing Habitat-rendered images. Numbers in gray indicated zero-shot inference results on held-out datasets.

**Fine-grain Language Understanding Benefits Target-oriented Navigation** We conduct multi-task training on DUET [19] initialized from LXMERT [99] using various data mixtures and report held-out inference results for datasets not included in the training. To address the visual gap between Habitat-rendered images and Matterport3D-rendered images, we map the R2R and REVERIE trajectories onto $\mathcal{G}^*$ and employ Habitat-rendered images for this experiment. The results are presented in Table 1.

Our findings reveal several key insights: (1) Mixing training data across different tasks reduces performance compared to training on individual tasks requiring higher-level language understanding capacity. For example, in the R2R task, incorporating additional data results in a 2-3% drop in success rate (SR), while in REVERIE, combining OBJECTNAV data leads to a substantial SR decrease of 6-7%, even lower than the zero-shot performance achieved when training exclusively with R2R data (39% SR vs. 35% SR). (2) Training with fine-grained human-annotated instruction-trajectory pairs proves advantageous for OBJECTNAV, yielding a 2-4% SR improvement. (3) Models trained exclusively on VLN data achieve strong zero-shot performance on OBJECTNAV (above 40% SR); however, models trained only on OBJECTNAV data perform poorly on tasks that demand sophisticated language understanding (with only $\sim$ 15% SR). Additionally, models trained solely on R2R (fine-grained VLN) data achieve 39% SR on REVERIE (coarse-grained VLN) and 55% SR on OBJECTNAV. This suggests that when the target is visible or only minor exploration is needed, R2R-trained models can successfully infer the location. These observations lead to two main findings:

1. **Fine-grained language understanding improves target-oriented navigation (OBJECTNAV)**, as visual-semantic understanding is enhanced through learning vision-language alignment; however, models trained exclusively on target-oriented data lack the language comprehension required to follow complex instructions.
2. **Training with simple data mixing is insufficient to achieve optimal performance** for tasks demanding both exploration and detailed instruction interpretation (coarse-grained VLN).

## 3. Mixture of Experts for Versatile Language-guided Visual Navigation

The insights from Section 2.2 motivate the need for a method to manage conflicts that arise during multi-task learning. To address this, we propose a new State-Adaptive Mixture of Experts (SAME) approach. SAME employs multiple specialized expert networks $\{f_1, \cdots, f_N\}$, which could be switched during each step in a navigation episode conditioned on the agent's state by a routing mechanism $\mathcal{R}$. In this way, we differentiate the learning of distinct navigation skills, such as exploration and instruction-following, while facilitating the sharing of common navigational knowledge like visual semantic understanding.

### 3.1. MoE Formulation

A subset of experts is activated in a sparsely gated MoE layer during each forward pass. The router predicts the probability of each expert being assigned:

$$\mathcal{P}(\mathbf{x}_r) = \text{Softmax}\left(\mathcal{R}\left(\mathbf{x}_r\right)\right), \tag{4}$$

$$\mathcal{R}(\mathbf{x}_r) = W\mathbf{x}_r, \tag{5}$$

where $\mathbf{x}_r$ is the routing feature extracted from the input $\mathbf{x}$, $W \in \mathbb{R}^{N\times d}$ is a trainable layer, $d$ is the hidden dimension and $N$ is the number of experts. The weighted sum of the outputs from experts with top-$k$ routing scores noted as set $\mathcal{T}$ is computed as the output:

$$\text{MoE}(\mathbf{x}, \mathbf{x}_r) = \sum_{i\in\mathcal{T}} \mathcal{P}\left(\mathbf{x}_r\right)_i f_i(\mathbf{x}). \tag{6}$$

Typically, a load balancing loss [30] is implemented to encourage an even distribution of input across the $N$ experts:

$$\mathcal{L}_{\text{balance}} = N \sum_{i=0}^{N} \mathcal{F}_i \mathcal{D}_i, \tag{7}$$

$$\mathcal{F} = \frac{1}{K} \sum_{i=1}^{N} \mathbf{1}\left\{\arg\max \mathcal{P}(\mathbf{x}_r) = i\right\}, \tag{8}$$

$$\mathcal{D} = \frac{1}{K} \sum_{i=1}^{K} \mathcal{P}(\mathbf{x}_r)_i, \tag{9}$$

where $\mathcal{F}$ represents the fraction of inputs processed by each top-$K$ expert $f_i$, and $\mathcal{D}$ represents the fraction of router probability allocated for expert $f_i$.

**Task-wised MoE and Token-wised MoE** Recently, most research on Mixture of Experts (MoE) has focused on transformer-based Large Language Models (LLMs), where MoE operates at the token level to process each individual input token, as illustrated in Figure 2. Concurrently, MoE is also extensively explored in computer vision multi-task learning, where expert modules are routed at the task level. These two routing features are formulated as follows:

$$\mathbf{x}_r^{\text{token}} = \mathbf{x}_i, \tag{10}$$

$$\mathbf{x}_r^{\text{task}} = E^{\text{task}} = W_t T, \tag{11}$$

where $\mathbf{x}_i$ is the $i$-th token in input $\mathbf{x}$, and $E^{\text{task}}$ is the task embedding for task with index $T$.

Beyond assigning each task $T$ to specific experts through a hard assignment, it might be helpful for an agent to select appropriate navigation skills based on the language instruction. To this end, we formulate a language-aware expert selection mechanism, represented as follows:

$$\mathbf{x}_r^{\text{text}} = \hat{\mathcal{W}}^{\texttt{CLS}}, \tag{12}$$

where $\hat{\mathcal{W}}^{\texttt{CLS}}$ denotes the $[\texttt{CLS}]$ token for text feature $\hat{\mathcal{W}}$.

### 3.2. State-Adaptive Experts Selection

Our initial experiments with the above token-wise and task-wise MoE in multi-task navigation learning yielded suboptimal results (to be presented in §3.3), prompting us to reconsider a more feasible MoE formulation for the sequential decision-making process in navigation. We observed that the density of language information the agent receives – which must align with visual observations to determine an action – can vary significantly across different timesteps in different tasks. In other words, the agent's state interpretation should be inherently generalizable to address distinct navigation problems. In light of this, we introduce SAME, a multimodal State-Adaptive expert selection mechanism:

$$\mathbf{x}_r^{\text{multi}} = W_m \left[\frac{1}{L_v}\sum_{i=0}^{L_v} \hat{\mathcal{O}}_t; \hat{\mathcal{W}}^{\texttt{CLS}}\right], \tag{13}$$

where a linear layer $W_m \in \mathbb{R}^{h\times 2h}$ is implemented to merge the concatenation of the mean visual feature $\frac{1}{L}\sum_{i=0}^{L} \hat{\mathcal{O}}_t$ over different views and the text $[\texttt{CLS}]$ token $\hat{\mathcal{W}}^{\texttt{CLS}}$.

Besides, we also investigate the effect of adding task information for expert selection:

$$\mathbf{x}_r^{\text{text\_task}} = \mathbf{x}_r^{\text{text}} + E^{\text{task}}, \tag{14}$$

$$\mathbf{x}_r^{\text{multi\_task}} = \mathbf{x}_r^{\text{multi}} + E^{\text{task}}. \tag{15}$$

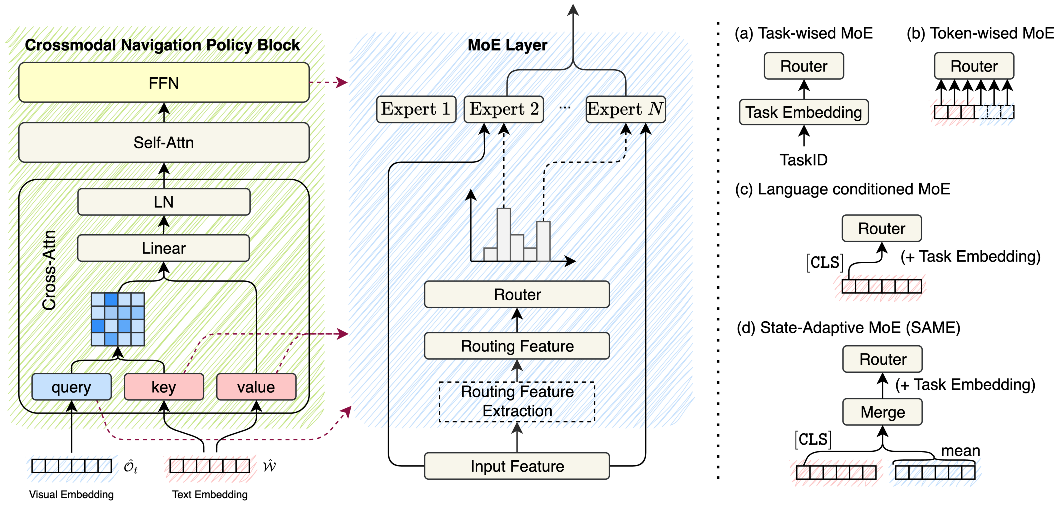


Figure 2. Illustration of MoE position and experts' routing methods. SAME routing based on multimodal features from visual observations and language instructions allows the agent to dynamically adapt to environmental visual changes.

| Routing Condition | R2R (Val Unseen) | | REVERIE (Val Unseen) | | ObjectNav-MP3D (Val) | |
|---|---|---|---|---|---|---|
| | SR↑ | SPL↑ | SR↑ | SPL↑ | SR↑ | SPL↑ |
| w/o MoE | $73.10_{\pm 0.29}$ | $64.93_{\pm 0.11}$ | $40.81_{\pm 2.54}$ | $33.41_{\pm 1.79}$ | $71.16_{\pm 2.65}$ | $43.89_{\pm 1.12}$ |
| *Token-wised MoE* [95]: | | | | | | |
| Token Embedding | $\mathbf{73.88}_{\pm 1.11}$ | $65.13_{\pm 1.21}$ | $42.02_{\pm 1.47}$ | $34.93_{\pm 1.08}$ | $72.73_{\pm 1.97}$ | $44.02_{\pm 1.59}$ |
| *Task-wised MoE*: | | | | | | |
| Task Embedding | $71.10_{\pm 0.39}$ | $64.73_{\pm 0.20}$ | $42.41_{\pm 2.14}$ | $34.70_{\pm 1.92}$ | $72.06_{\pm 2.15}$ | $41.39_{\pm 3.09}$ |
| Text `[CLS]` | $73.41_{\pm 0.37}$ | $64.23_{\pm 0.21}$ | $43.24_{\pm 2.15}$ | $34.65_{\pm 1.27}$ | $72.56_{\pm 2.64}$ | $43.52_{\pm 1.62}$ |
| w/ Task Embedding | $72.93_{\pm 0.45}$ | $62.67_{\pm 0.82}$ | $42.59_{\pm 2.35}$ | $33.40_{\pm 1.97}$ | $71.77_{\pm 2.31}$ | $41.91_{\pm 3.13}$ |
| *State-Adaptive MoE (ours)*: | | | | | | |
| SAME | $73.86_{\pm 0.36}$ | $\mathbf{65.19}_{\pm 0.71}$ | $\mathbf{43.84}_{\pm 2.15}$ | $\mathbf{35.30}_{\pm 2.07}$ | $\mathbf{72.87}_{\pm 2.43}$ | $43.92_{\pm 1.03}$ |
| w/ Task Embedding | $73.46_{\pm 0.44}$ | $64.81_{\pm 0.26}$ | $41.85_{\pm 3.40}$ | $34.68_{\pm 2.68}$ | $70.90_{\pm 2.17}$ | $\mathbf{44.59}_{\pm 1.45}$ |
| w/o Pretrain | $59.89_{\pm 0.37}$ | $49.73_{\pm 0.91}$ | $38.14_{\pm 1.15}$ | $32.10_{\pm 0.97}$ | $69.76_{\pm 2.47}$ | $38.24_{\pm 2.22}$ |

Table 2. Comparison of performance with different MoE routing conditions on R2R, REVERIE, and ObjectNav-MP3D.

### 3.3. Comparison on MoE Routing

Following the above, we can see that the key to distinguishing specific and shared navigational knowledge learning is learning the routing mechanism in MoE. In this section, we investigate the routing strategies for a versatile language-guided navigation policy.

**Experiment Setup** We conduct contrastive experiments to compare the routing mechanism. Matterport3D-rendered images [6] are used for R2R and REVERIE. We adopt 8 experts for each MoE layer and choose the top 2 experts [30]. The MoE experts are deployed to the visual queries (will be discussed in Section 3.4). We report the statistic results from 5 checkpoints after training converged, shown in Table 2. As discussed in Section 2.2, we demonstrate that learning fine-grained language understanding can enhance target-oriented navigation learning, motivating us to take advantage of the powerful vision-language-action pre-train conducted in VLN [17, 19, 35, 37, 75, 82, 108]. Throughout the subsequent experiments, we initialized our model from ScaleVLN [108], which scales the VLN data from 14K to 4.9M by generating synthetic data to perform pertaining.

**VLN Pretrain Benefits Navigation Multi-task Learning** As shown in Table 2, there is a significant improvement on all tasks when initializing with VLN pre-trained weights compared to directly performing multi-task tuning on SAME initialized from general vision-language pretrain LXMERT [99] (w/o Pretrain), featured by a $\sim 14\%$ SR increase on R2R.

**MoE Benefits Navigation Multi-task Learning** Compared to the multi-task-tuned DUET (w/o MoE) under the same experimental conditions, all MoE methods show a significant performance increase (2-3% SR) on the REVERIE task, with no notable performance drops on other tasks. This aligns with our findings in Section 2.2, suggesting that coarse-grained VLN performance is influenced when co-trained with other language instruction types, likely due to model overfitting to the R2R task. Our proposed MoE approach addresses this issue effectively.

**SAME Facilitates Navigation Skill Sharing** Comparing different MoE routing types, we highlight a significant improvement of $\sim$ 2% SR and $\sim$ 1% SPL in SAME routing on REVERIE. We hypothesize that this improvement is due to the nature of the coarse-grained VLN task, where the agent must alternate between exploration, in cases where language guidance lacks detail, and adhering closely to language instructions when precisely localizing the target. SAME routing enables a flexible selection of experts to manage these distinct navigation behaviors based on the current observation and language input, allowing the agent to learn transferable knowledge across tasks. This flexibility also results in the SAME agent achieving the highest average SPL of 48.13% across all tasks.

**Explicit Task Information Hinders Skill Sharing** Furthermore, we examine hard assignments with specific ex-

| MoE Experts Position | R2R (Val Unseen) | | REVERIE (Val Unseen) | | ObjectNav-MP3D (Val) | |
|---|---|---|---|---|---|---|
| | SR↑ | SPL↑ | SR↑ | SPL↑ | SR↑ | SPL↑ |
| Feed Forward | $73.27_{\pm 0.47}$ | $63.74_{\pm 0.62}$ | $43.42_{\pm 2.01}$ | $35.16_{\pm 1.93}$ | $71.93_{\pm 2.05}$ | $43.08_{\pm 1.75}$ |
| Visual Query | $\mathbf{73.86}_{\pm 0.36}$ | $\mathbf{65.19}_{\pm 0.71}$ | $\mathbf{43.84}_{\pm 2.15}$ | $\mathbf{35.30}_{\pm 2.07}$ | $\mathbf{72.87}_{\pm 2.43}$ | $43.92_{\pm 1.03}$ |
| Textual Key & Value | $73.35_{\pm 0.32}$ | $63.98_{\pm 0.92}$ | $43.32_{\pm 1.97}$ | $34.78_{\pm 1.86}$ | $72.35_{\pm 2.02}$ | $\mathbf{44.08}_{\pm 0.95}$ |

Table 3. Comparison of performance with different MoE experts' positions on R2R, REVERIE, and ObjectNav-MP3D.

perts dedicated to different tasks by directly routing through task embeddings. This approach results in performance drops across all tasks. Additionally, incorporating task embeddings into $\mathbf{x}_r^{\text{text}}$ and $\mathbf{x}_r^{\text{multi}}$ reduces performance, indicating that routing based on samples from different tasks provides strong but ineffective bias, preventing the router from effectively learning which experts to select based on the instruction or observation, thus impeding the development of shared knowledge across tasks.

### 3.4. Which Part of the Navigation Policy Learns Different Navigation Behaviour?

Since the initial application of MoE in transformer architectures [30, 56, 131], MoE has acted as an enhancement for Feed-Forward Network (FFN) modules within these models. Concurrently, some studies have integrated multi-head attention layers with MoE to enhance performance further while managing computational costs. In this section, we analyze the impact of MoE applied at different components of the transformer model, specifically focusing on the FFN, visual queries $W_q$, textual key $W_k$, and value $W_v$ with SAME under the same experiment setup described in Section 3.3. The results are shown in Table 3.

MoE applied to different components yields varying performance across tasks. Notably, the best overall performance is observed when applying MoE to the visual query, achieving the highest SR on all tasks with fewer parameters. This suggests that utilizing MoE at the visual query level within the cross-attention layer is particularly effective. This effectiveness likely stems from the multimodal policy's control over diverse navigation behaviors within the cross-attention layer, where specialized visual query experts allow the agent to more accurately determine the next action by adjusting attention scores over visual embeddings from multiple viewpoints. While FFN-based MoE enhances performance compared to non-MoE models, it is surpassed by MoE configurations that integrate with attention layers, highlighting the crucial role of cross-modal attention in successful action selection.

In summary, the experimental results indicate that employing MoE with visual query experts and routing based on multimodal features from visual observations and language instructions (SAME) allows the agent to dynamically adapt to environmental visual changes while staying aligned with language guidance, thereby enhancing robust performance across different language-guided navigation tasks.

## 4. Experiments

In this section, we conduct larger scale multi-task tuning based on the previous discussion. We conduct a thorough evaluation of SAME across seven major navigation benchmarks, complemented by a series of ablation studies on training details to establish best practices for effective multi-task training in language-guided navigation.

**Datasets** Besides the R2R, REVERIE, and ObjectNav-MP3D, we include 4 other datasets, RxR-EN [54], CVDN [101], SOON [129], and R2R-CE [53], for larger scale training and evaluation. We leave the detailed introduction to these datasets in the supplementary.

**Implementation Details** We build upon the DUET architecture and replace all the visual query layers in the cross-attention with MoE layers. We adopt SAME routing for each MoE layer. We initialize from the pre-trained weights of ScaleVLN [108] and utilize CLIP ViT-B/16 [86] as the visual encoder. We use ScaleVLN, R2R, RxR-EN, CVDN, REVERIE, SOON, and Habitat-Web as the training data with a sampling ratio of 10:1:1:1:1:1:2 without mixing different data in a batch. We follow [3, 19, 48] to utilize DAgger [92] algorithm to obtain interactive supervision from the simulator. The training objective $\mathcal{L} = \mathcal{L}_{\text{DAG}} + \lambda \mathcal{L}_{\text{balance}}$ is the combination of DAgger loss and MoE load balancing loss in Equation 7, balanced by coefficient $\lambda = 0.8$. The model is fine-tuned using AdamW optimizer [72] with a learning of $1 \times 10^{-5}$ for 20k iterations with a batch size of 16 on a single 80G NVIDIA A100 GPU.

### 4.1. Comparison with State-of-the-Art Models

**Comparison in Discrete Environment** We first compare our method with current SoTA methods in the discrete MP3D environment [6]. The datasets are organized in Table 4 by language instruction granularity and complexity, ranging from fine-grained and complex to coarse-grained and simple, from left to right. We classify previous methods into two categories. The first one is a separate model for each task, which fine-tunes a distinct set of parameters for each downstream task. The second one is a unified model for all tasks, which includes our baseline methods:

- MT-RCM [106]: This method performs multi-task training on R2R and CVDN using the RCM model [105].
- NaviLLM [125]: A hybrid model of DUET and Large Language Model (LLM) that employs a stronger vision encoder, EVA-CLIP-02-Large [29], and replaces the 12-layer cross-modal encoder in DUET with an LLM, Vicuna-7B [23], to facilitate multi-task instruction tuning.
- ScaleVLN†: We conduct multi-task tuning on DUET [19] with ScaleVLN [108] initilization. This serves as a baseline for multi-task tuning on DUET without using MoE.

As shown in Table 4, SAME achieve State-of-the-Art performance on multiple benchmarks with a unified model.

| Methods | CVDN | | RxR-EN | | R2R | | | | SOON | | | | REVERIE | | | | ObjectNav-MP3D | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Val | Test | Val unseen | | Val unseen | | Test unseen | | Val unseen | | Test unseen | | Val unseen | | Test unseen | | Val | |
| | GP ↑ | GP ↑ | SR ↑ | nDTW ↑ | SR ↑ | SPL ↑ | SR ↑ | SPL ↑ | SR ↑ | SPL ↑ | SR ↑ | SPL ↑ | SR ↑ | SPL ↑ | SR ↑ | SPL ↑ | SR ↑ | SPL ↑ |
| *Separate Model for Each Task*: | | | | | | | | | | | | | | | | | | |
| SF [31] | - | - | - | - | 36 | - | 35 | 28 | - | - | - | - | - | - | - | - | - | - |
| RCM [105] | - | - | - | - | 43 | - | 43 | 38 | - | - | - | - | 9.3 | 7.0 | 7.8 | 6.7 | - | - |
| EnvDrop [100] | - | - | - | - | 52 | 48 | 51 | 47 | - | - | - | - | - | - | - | - | - | - |
| PREVALENT [37] | 3.15 | 2.44 | - | - | 58 | 53 | 54 | 51 | - | - | - | - | - | - | - | - | - | - |
| VLN↻BERT [40] | - | - | - | - | 63 | 57 | 63 | 57 | - | - | - | - | 25.5 | 21.1 | 24.6 | 19.5 | - | - |
| HAMT [18] | 5.13 | 5.58 | 56.4 | 63.0 | 66 | 61 | 65 | 60 | - | - | - | - | 33.0 | 30.2 | 30.4 | 26.7 | - | - |
| HOP+ [82] | - | - | - | - | 67 | 61 | 66 | 60 | - | - | - | - | 36.1 | 31.1 | 33.8 | 28.2 | - | - |
| DUET [19] | - | - | - | - | 72 | 60 | 69 | 59 | 36.3 | 22.6 | 33.4 | 21.4 | 47.0 | 33.7 | 52.5 | 36.1 | - | - |
| AutoVLN [20] | - | - | - | - | - | - | - | - | 41.0 | 30.7 | 40.4 | 27.9 | 55.9 | 40.9 | 55.2 | 38.9 | - | - |
| BEVBert [3] | - | - | 66.7 | 69.6 | 75 | 64 | 73 | 62 | - | - | - | - | 51.8 | 36.4 | 52.8 | 36.4 | - | - |
| GridMM [109] | - | - | - | - | 75 | 64 | 73 | 62 | - | - | - | - | - | - | - | - | - | - |
| VER [68] | - | | - | - | 76 | 65 | 76 | 66 | - | - | - | - | 56.0 | 39.7 | 56.8 | 38.8 | - | - |
| GOAT [103] | - | - | 68.2 | 66.8 | 78 | 68 | 75 | 65 | 40.4 | 28.1 | 40.5 | 25.2 | 53.4 | 36.7 | 57.7 | 40.5 | - | - |
| ScaleVLN [108] | 6.12 | 6.97 | - | - | 79 | 70 | 77 | 68 | - | - | - | - | 57.0 | 41.8 | 56.1 | 39.5 | - | - |
| *Unified Model for All Tasks*: | | | | | | | | | | | | | | | | | | |
| MT-RCM [106] | 4.65 | 3.91 | - | - | 47 | 41 | 45 | 40 | - | - | - | - | - | - | - | - | - | - |
| NaviLLM [125] | 6.16 | **7.90** | - | - | 67 | 59 | 68 | 60 | **38.3** | **29.2** | 35.0 | 26.3 | 42.2 | 35.7 | 39.8 | 32.3 | - | - |
| ScaleVLN† | 5.93 | - | 46.7 | 49.7 | **76** | **67** | - | - | 33.2 | 25.4 | - | - | 41.9 | 34.4 | - | - | 72.3 | **43.4** |
| SAME (ours) | **6.94** | 7.07 | **50.5** | **51.2** | **76** | 66 | **74** | **64** | 36.1 | 25.4 | **38.2** | **27.1** | **46.4** | **36.1** | **48.6** | **37.1** | **76.3** | 42.7 |

Table 4. Agents performance across all tasks in the discrete environment [6]. † indicates our implementation of multi-task tuning. Note that existing methods tailored for ObjectNav-MP3D are proposed in continuous environments, which will be evaluated in Table 5 below.

| Methods | R2R-CE (Val unseen) | | | MP3D (Val) | |
|---|---|---|---|---|---|
| | NE ↓ | SR ↑ | SPL↑ | SR ↑ | SPL ↑ |
| NaVid [120] | 5.47 | 37 | 36 | - | - |
| ScaleVLN [108] | 4.80 | 55 | 51 | - | - |
| ETPNav [4] | 4.71 | 57 | 49 | - | - |
| BEVBert [3] | **4.57** | **59** | **50** | - | - |
| SemExp [10] | - | - | - | 28 | 11 |
| PONI [88] | - | - | - | 32 | 12 |
| Habitat-Web [89] | - | - | - | 35 | 10 |
| SAME (ours) | 5.31 | 47 | 38 | **43** | **21** |

Table 5. Comparison with previous methods in the continuous environment [94]. We report the zero-shot inference results of SAME using the same checkpoint from Table 4.

Compared to separate models for each task, SAME achieve SoTA performance on CVDN with a 7.07 GP on test 13% GP increase on validation compared to ScaleVLN. SAME performs at the same level as VER and GOAT, comparing the SPL on R2R and REVERIE. Compared to other unified models,  outperforms the baseline multi-task tuned ScaleVLN on all tasks, with an average of 3% SR increase among all tasks. SAME performs significantly better than NaviLLM on R2R and REVERIE (6% and 9% higher test SR) with much fewer parameters (Table 6).

**Comparison in Continuous Environment** We further compare our method with current SoTA methods in the continuous Habitat environment [94]. We take the same model reported in Table 4 and examine the zero-shot inference results on held-out datasets in continuous environment. We follow Hong *et al.* [41] and deploy the waypoint predictor to bridge the gap between discrete and continuous. As shown in Table 5, SAME surpass Habitat-Web, which performs imitation learning on ObjectNav human demonstration, when zero-shot transferring to the continuous environment. This aligned with our findings in §2.2 that fine-grain language understanding improves target-oriented navigation.

| Model | # Params | MACs | FLOPs | Latency |
|---|---|---|---|---|
| DUET | 180.29 M | 17.75 G | 35.53 G | 28.78 ms |
| SAME (ours) | 215.74 M | 18.68 G | 37.38 G | 59.14 ms |
| NaviLLM | 6.77 B | 2.11 T | 4.23 T | 99.87 ms |

Table 6. Comparison of computational overhead.

## 4.2. Computational Overhead

We provide a detailed analysis of computational overhead introduced by SAME in Table 6. Compared to DUET, 34.45M more parameters are introduced by SAME, but only 5.2% more FLOPs are increased, as the two models share 83.6% of parameters and only top-2 experts are activated at each step. SAME consists of significantly fewer parameters than NaviLLM and achieves much better performance.

## 4.3. Ablation Studies

We conduct a series of experiments to establish best practices for training with multiple navigation tasks. These ablations are performed using the same experimental setup as described in Section 3.3.

**Effect of Training Schema** We investigate different training schemas for SAME, specifically the training algorithm and data mixing strategy. For the training algorithm, we compare training with imitation learning only on teacher actions [6] to training with DAgger, where at each time step, an agent performs an action sampled from the predicted

| Algorithm | Batch | R2R (Val Unseen) | | | | REVERIE (Val Unseen) | | | | ObjectNav-MP3D (Val) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | TL | NE↓ | SR↑ | SPL↑ | TL | NE↓ | SR↑ | SPL↑ | TL | NE↓ | SR↑ | SPL↑ |
| Imitation | Mix | 13.77 | 3.82 | 65.51 | 56.57 | 19.02 | 5.79 | 35.64 | 28.15 | 31.31 | 2.64 | **75.83** | 26.77 |
| Imitation | Sequential | 9.35 | 4.23 | 61.09 | 58.88 | 9.24 | 7.37 | 28.37 | 26.42 | 26.57 | 3.51 | 71.70 | 24.06 |
| DAgger | Mix | 16.09 | 4.07 | 62.36 | 52.18 | 25.05 | 5.89 | 30.45 | 21.91 | 27.57 | 2.81 | 74.85 | 31.70 |
| DAgger | Sequential | 13.51 | 2.90 | **73.69** | **64.92** | 16.32 | 5.38 | **45.67** | **37.95** | 15.60 | 3.10 | 71.43 | **43.39** |

Table 7. Comparison of single-run performance with different training schema on R2R, REVERIE, and ObjectNav-MP3D.

| $\lambda$ | R2R (Val Unseen) | | | | REVERIE (Val Unseen) | | | | ObjectNav-MP3D (Val) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TL | NE↓ | SR↑ | SPL↑ | TL | NE↓ | SR↑ | SPL↑ | TL | NE↓ | SR↑ | SPL↑ |
| 0.2 | 12.63 | 2.91 | 73.61 | **65.73** | 15.47 | 5.55 | 42.86 | 35.22 | 15.25 | 3.13 | 71.79 | 43.51 |
| 0.5 | 13.77 | 2.89 | **74.29** | 65.00 | 19.45 | 5.56 | 44.70 | 34.89 | 18.10 | 3.05 | 72.26 | 40.50 |
| 0.8 | 13.51 | 2.90 | 73.69 | 64.92 | 16.32 | 5.38 | 45.67 | **37.95** | 15.60 | 3.10 | 71.43 | 43.39 |
| 1.0 | 13.56 | 3.02 | 73.56 | 64.52 | 16.19 | 5.24 | **45.81** | 37.89 | 18.11 | 2.67 | **74.66** | **44.81** |

Table 8. Comparison of single-run performance with different MoE balance coefficients $\lambda$ on R2R, REVERIE, and ObjectNav-MP3D.

probability of its action space and minimizes the loss between the sampled action and the ground truth. This method allows an agent to learn from paths that cover wide space and reduces the exposure bias caused by teacher forcing [46]. For the data mixing strategy, we investigate mixing different datasets in the same batch versus sampling different data for training. As shown in Table 7, training with DAgger significantly improves the performance when sequentially sampling data. Compare row 4 to row 2 in Table 7, 12.6% SR and 6.04% SPL increase on R2R, 17.3% SR and 11.53% SPL increase on REVERIE, 19.33% SPL increase on ObjectNav-MP3D is observed. We also noticed that mixing up training data in the same batch would significantly affect DAgger training, comparing row 4 to row 3. This indicates utilizing Dagger without mixing up training data in the same batch is the best practice.

**Effect of MoE Routing Balance Coefficient** We investigate the effect of the MoE routing balance loss coefficient $\lambda$ in Table 8. When $\lambda$ is large, the balancing loss would force the model to select different experts for samples in the same batch evenly; when $\lambda$ decreases, such constraint is weakened. The model trained with $\lambda = 0.2$ in row 1 performs worse than all other variants, demonstrating the significance of the balancing loss in MoE training.

# 5. Related Work

## 5.1. Vision-and-Language Navigation

The development of a navigation agent capable of interpreting and acting upon unrestricted linguistic instructions to navigate through unfamiliar, photorealistic environments has been a longstanding goal within the field of Vision-and-Language Navigation [7, 39, 54, 81, 101, 122, 129]. Approaches to this problem have primarily addressed two main areas: (1) Vision-Language Alignment: Some studies [17, 35, 37, 40, 59, 62, 75, 82, 123] leverage generic visual-linguistic representations [21, 61, 63, 98, 99]. Others incorporate additional supervision through data augmentation [5, 27, 35, 57, 58, 60, 80, 100, 108], along with training strategies [43, 73, 104, 105, 128] to enhance cross-modal alignment. (2) Efficient Action Planning Mechanisms: These involve historical state memorization [18, 40], self-correction strategies [38, 49, 74], navigation map construction [2, 14, 19, 67, 68, 102, 109, 124]. Recently, some work [13, 65, 70, 71, 79, 83, 85, 119–121, 125–127] integrate LLMs for generalizable world knowledge.

## 5.2. ObjectGoal Navigation

Approaches to learning to understand visual semantics and perform ObjectNav [9, 25, 26, 28, 32, 50–52, 87, 113, 118] could be categorized into two streams: (1) Modular Pipelines with Learned Modules [10, 11, 36, 88, 89, 111]: This paradigm integrates learning into specific modules by leveraging explicit scene representations like semantic map [11, 36, 88] or simply employing object detectors or segmentors [76, 89]. (2) End-to-end Learning with RL or IL [1, 90, 112, 115, 117]: these methods benefit from visual representation [77, 116], auxiliary task [117], and data augmentation [76] to generalize to unseen environments.

## 5.3. Mixture of Experts

Mixture of Experts (MoE) [45, 47, 95] models utilize multiple specialized experts along with a routing network that dynamically assigns tasks based on their complexity. Task-oriented MoE enhances the model's capacity to learn both specific and shared knowledge in computer vision, without significantly increasing computational costs [22, 34, 69, 78, 91, 96]. Sparsely activated MoE is widely employed in LLMs to reduce computational costs while enabling the training of gigantic models with trillions of parameters through sparse activation [30, 56, 95]. MoE assigns different experts for instruction tuning in recent LLMs [24, 46, 114], optimizing the learning process by focusing on specific tasks within the overall model architecture. To address task conflicts and enhance generalization in unseen tasks, some methods [22, 34, 107] employ various strategies to optimize the selection and aggregation of expert outputs in MoE.

# 6. Conclusion

This paper unifies a diverse range of navigation tasks within a cohesive language-guided navigation framework. It examines the fundamental challenges of sharing common knowledge while leveraging task-specific capabilities in navigation learning. We propose the State-Adaptive Mixture of Experts (SAME), which enables an agent to make decisions by integrating multi-granularity language inputs and dynamic observations. We believe SAME can guide the learning towards versatile language-guided navigation agents.

## Acknowledgements

This work was supported by the Centre for Augmented Reasoning, an initiative by the Department of Education, Australian Government.

## References

[1] Ziad Al-Halah, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Zero experience required: Plug & play modular transfer learning for semantic visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17031–17041, 2022. 8

[2] Dong An, Yuankai Qi, Yan Huang, Qi Wu, Liang Wang, and Tieniu Tan. Neighbor-view enhanced model for vision and language navigation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5101–5109, 2021. 8

[3] Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. Bevbert: Topo-metric map pre-training for language-guided navigation. *arXiv preprint arXiv:2212.04385*, 2022. 6, 7

[4] Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *arXiv preprint arXiv:2304.03047*, 2023. 7

[5] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 3, 8

[6] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018. 1, 2, 3, 5, 6, 7

[7] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 8

[8] 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. 1, 2, 3

[9] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, 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)*, pages 667–676. IEEE, 2017. 3, 8

[10] 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. 1, 7, 8

[11] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12875–12884, 2020. 1, 8

[12] Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. *arXiv preprint arXiv:2407.05890*, 2024. 1

[13] Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee K Wong. Mapgpt: Map-guided prompting for unified vision-and-language navigation. *arXiv preprint arXiv:2401.07314*, 2024. 1, 8

[14] Kevin Chen, Junshen K Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. Topological planning with transformers for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11276–11286, 2021. 8

[15] Kehan Chen, Dong An, Yan Huang, Rongtao Xu, Yifei Su, Yonggen Ling, Ian Reid, and Liang Wang. Constraint-aware zero-shot vision-language navigation in continuous environments. *arXiv preprint arXiv:2412.10137*, 2024. 1

[16] Peihao Chen, Xinyu Sun, Hongyan Zhi, Runhao Zeng, Thomas H Li, Gaowen Liu, Mingkui Tan, and Chuang Gan. $a^2$ nav: Action-aware zero-shot robot navigation by exploiting vision-and-language ability of foundation models. *arXiv preprint arXiv:2308.07997*, 2023. 1

[17] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34:5834–5847, 2021. 1, 5, 8

[18] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34:5834–5847, 2021. 7, 8

[19] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16537–16547, 2022. 1, 2, 3, 5, 6, 7, 8

[20] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Learning from unlabeled 3d environments for vision-and-language navigation. In *European Conference on Computer Vision*, pages 638–655. Springer, 2022. 7

[21] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1, 8

[22] Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, Yu Qiao, and Jing Shao. Octavius: Mitigating task interference in mllms via moe. *arXiv preprint arXiv:2311.02684*, 3, 2023. 8

[23] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 6

[24] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024. 1, 8

[25] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3164–3174, 2020. 8

[26] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35: 5982–5994, 2022. 8

[27] Zi-Yi Dou and Nanyun Peng. Foam: A follower-aware speaker model for vision-and-language navigation. *arXiv preprint arXiv:2206.04294*, 2022. 8

[28] Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Manipulathor: A framework for visual object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4497–4506, 2021. 8

[29] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 6

[30] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 4, 5, 6, 8

[31] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 7

[32] SY Gadre, M Wortsman, G Ilharco, L Schmidt, and S Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. *arXiv preprint arXiv:2203.10421*, 2022. 8

[33] Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3064–3073, 2021. 1

[34] Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023. 8

[35] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1634–1643, 2021. 1, 5, 8

[36] Meera Hahn, Devendra Singh Chaplot, Shubham Tulsiani, Mustafa Mukadam, James M Rehg, and Abhinav Gupta. No rl, no simulation: Learning to navigate without navigating. *Advances in Neural Information Processing Systems*, 34:26661–26673, 2021. 8

[37] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146, 2020. 1, 5, 7, 8

[38] Haodong Hong, Sen Wang, Zi Huang, Qi Wu, and Jiajun Liu. Navigating beyond instructions: Vision-and-language navigation in obstructed environments. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7639–7648, 2024. 8

[39] Yicong Hong. *Learning Language-Guided Visual Navigation*. PhD thesis, The Australian National University (Australia), 2023. 8

[40] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653, 2021. 1, 7, 8

[41] Yicong Hong, Zun Wang, Qi Wu, and Stephen Gould. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15439–15449, 2022. 3, 7

[42] Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6551–6557, 2019. 1

[43] Haoshuo Huang, Vihan Jain, Harsh Mehta, Alexander Ku, Gabriel Magalhaes, Jason Baldridge, and Eugene Ie. Transferable representation learning in vision-and-language navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7404–7413, 2019. 8

[44] Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. *arXiv preprint arXiv:1907.05446*, 2019. 3

[45] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 1, 8

[46] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 1, 8

[47] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994. 1, 8

[48] Aishwarya Kamath, Peter Anderson, Su Wang, Jing Yu Koh, Alexander Ku, Austin Waters, Yinfei Yang, Jason Baldridge, and Zarana Parekh. A new path: Scaling vision-and-language navigation with synthetic instructions and imitation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10813–10823, 2023. 6

[49] Liyiming Ke, Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, and Siddhartha Srinivasa. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6741–6749, 2019. 8

[50] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393, 2024. 8

[51] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16373–16383, 2024.

[52] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 8

[53] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 2, 6

[54] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, 2020. 1, 6, 8

[55] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, 2020. 2

[56] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. 6, 8

[57] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *arXiv preprint arXiv:2305.19195*, 2023. 8

[58] Jialu Li and Mohit Bansal. Improving vision-and-language navigation by generating future-view image semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10803–10812, 2023. 8

[59] Jialu Li, Hao Tan, and Mohit Bansal. Improving cross-modal alignment in vision language navigation via syntactic information. *arXiv preprint arXiv:2104.09580*, 2021. 8

[60] Jialu Li, Hao Tan, and Mohit Bansal. Envedit: Environment editing for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15407–15417, 2022. 8

[61] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1, 8

[62] Xiujun Li, Chunyuan Li, Qiaolin Xia, Yonatan Bisk, Asli Celikyilmaz, Jianfeng Gao, Noah Smith, and Yejin Choi. Robust navigation with language pretraining and stochastic sampling. *arXiv preprint arXiv:1909.02244*, 2019. 1, 8

[63] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 8

[64] Xiwen Liang, Liang Ma, Shanshan Guo, Jianhua Han, Hang Xu, Shikui Ma, and Xiaodan Liang. Cornav: Autonomous agent with self-corrected planning for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2306.10322*, 2023. 1

[65] Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *arXiv preprint arXiv:2403.07376*, 2024. 8

[66] Xiangru Lin, Guanbin Li, and Yizhou Yu. Scene-intuitive agent for remote embodied visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7036–7045, 2021. 1

[67] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10968–10980, 2023. 8

[68] Rui Liu, Wenguan Wang, and Yi Yang. Volumetric environment representation for vision-language navigation. In

*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16317–16328, 2024. 7, 8

[69] Zhili Liu, Kai Chen, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, and James T Kwok. Task-customized masked autoencoder via mixture of cluster-conditional experts. *arXiv preprint arXiv:2402.05382*, 2024. 8

[70] Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. Discuss before moving: Visual language navigation via multi-expert discussions. *arXiv preprint arXiv:2309.11382*, 2023. 1, 8

[71] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *arXiv preprint arXiv:2406.04882*, 2024. 1, 8

[72] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6

[73] Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint arXiv:1901.03035*, 2019. 8

[74] Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. The regretful agent: Heuristic-aided navigation through progress estimation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 6732–6740, 2019. 8

[75] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*, pages 259–274. Springer, 2020. 1, 5, 8

[76] 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*, pages 15374–15383, 2021. 8

[77] Arsalan Mousavian, Alexander Toshev, Marek Fišer, Jana Košecká, Ayzaan Wahid, and James Davidson. Visual representations for semantic target driven navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8846–8852. IEEE, 2019. 8

[78] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022. 8

[79] Bowen Pan, Rameswar Panda, SouYoung Jin, Rogerio Feris, Aude Oliva, Phillip Isola, and Yoon Kim. Langnav: Language as a perceptual representation for navigation. *arXiv preprint arXiv:2310.07889*, 2023. 8

[80] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Javen Qinfeng Shi, and Anton van den Hengel. Counterfactual vision-and-language navigation: Unravelling the unseen. *Advances in Neural Information Processing Systems*, 33:5296–5307, 2020. 8

[81] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020. 1, 2, 3, 8

[82] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop+: History-enhanced and order-aware pre-training for vision-and-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 5, 7, 8

[83] Yanyuan Qiao, Yuankai Qi, Zheng Yu, Jing Liu, and Qi Wu. March in chat: Interactive prompting for remote embodied referring expression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15758–15767, 2023. 8

[84] Yanyuan Qiao, Wenqi Lyu, Hui Wang, Zixu Wang, Zerui Li, Yuan Zhang, Mingkui Tan, and Qi Wu. Open-nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source llms. *arXiv preprint arXiv:2409.18794*, 2024. 1

[85] Yanyuan Qiao, Qianyi Liu, Jiajun Liu, Jing Liu, and Qi Wu. Llm as copilot for coarse-grained vision-and-language navigation. In *European Conference on Computer Vision*, pages 459–476. Springer, 2025. 8

[86] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6

[87] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 8

[88] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18890–18900, 2022. 1, 7, 8

[89] Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5173–5183, 2022. 3, 7, 8

[90] 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. 8

[91] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel

Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 8

[92] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011. 6

[93] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. *ICCV*, 2019. 3

[94] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 3, 7

[95] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 1, 5, 8

[96] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*, 2023. 8

[97] Gunnar A Sigurdsson, Jesse Thomason, Gaurav S Sukhatme, and Robinson Piramuthu. Rrex-bot: Remote referring expressions with a bag of tricks. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5203–5210. IEEE, 2023. 1

[98] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1, 8

[99] 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)*, pages 5100–5111, 2019. 1, 3, 5, 8

[100] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of NAACL-HLT*, pages 2610–2621, 2019. 1, 7, 8

[101] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406, 2020. 1, 2, 6, 8

[102] Hanqing Wang, Wenguan Wang, Tianmin Shu, Wei Liang, and Jianbing Shen. Active visual information gathering for vision-language navigation. In *European Conference on Computer Vision*, pages 307–322. Springer, 2020. 8

[103] Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13139–13150, 2024. 7

[104] Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 37–53, 2018. 8

[105] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6629–6638, 2019. 1, 6, 7, 8

[106] Xin Eric Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. Environment-agnostic multitask learning for natural language grounded navigation. In *European Conference on Computer Vision*, pages 413–430. Springer, 2020. 6, 7

[107] Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. *arXiv preprint arXiv:2205.12410*, 2022. 8

[108] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12009–12020, 2023. 5, 6, 7, 8

[109] 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. 7, 8

[110] Zehao Wang, Minye Wu, Yixin Cao, Yubo Ma, Meiqi Chen, and Tinne Tuytelaars. Navigating the nuances: A fine-grained evaluation of vision-language navigation. *arXiv preprint arXiv:2409.17313*, 2024. 1

[111] Justin Wasserman, Karmesh Yadav, Girish Chowdhary, Abhinav Gupta, and Unnat Jain. Last-mile embodied visual navigation. In *Conference on Robot Learning*, pages 666–678. PMLR, 2023. 8

[112] 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 *International Conference on Learning Representations (ICLR)*, 2020. 8

[113] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9068–9079, 2018. 8

[114] Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*, 2024. 1, 8

[115] Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline visual representation learning for embodied navigation. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023. 8

[116] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. *arXiv preprint arXiv:1810.06543*, 2018. 8

[117] Joel Ye, Dhruv Batra, Abhishek Das, and Erik Wijmans. Auxiliary tasks and exploration enable objectgoal navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16117–16126, 2021. 8

[118] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, et al. Homerobot: Open-vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565*, 2023. 8

[119] Zhaohuan Zhan, Lisha Yu, Sijie Yu, and Guang Tan. Mc-gpt: Empowering vision-and-language navigation with memory map and reasoning chains. *arXiv preprint arXiv:2405.10620*, 2024. 8

[120] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and Wang He. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024. 7

[121] Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. *arXiv preprint arXiv:2502.13451*, 2025. 8

[122] Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035*, 2024. 8

[123] Chongyang Zhao, Yuankai Qi, and Qi Wu. Mind the gap: Improving success rate of vision-and-language navigation by revisiting oracle success routes. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4349–4358, 2023. 8

[124] Yusheng Zhao, Jinyu Chen, Chen Gao, Wenguan Wang, Lirong Yang, Haibing Ren, Huaxia Xia, and Si Liu. Target-driven structured transformer planner for vision-language navigation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4194–4203, 2022. 8

[125] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. *arXiv preprint arXiv:2312.02010*, 2023. 1, 6, 7, 8

[126] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7641–7649, 2024. 1

[127] Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, pages 260–278. Springer, 2025. 8

[128] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10012–10022, 2020. 8

[129] Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12689–12699, 2021. 1, 2, 6, 8

[130] 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)*, pages 3357–3364. IEEE, 2017. 1

[131] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022. 6