

Embodied Navigation with Auxiliary Task of Action Description Prediction

Haru Kondoh¹ Asako Kanezaki^{1,2}

¹ Institute of Science Tokyo ² RIKEN AIP

kondo.h.4aa3@m.isct.ac.jp kanezaki@comp.isct.ac.jp

Abstract

The field of multimodal robot navigation in indoor environments has garnered significant attention in recent years. However, as tasks and methods become more advanced, the action decision systems tend to become more complex and operate as black-boxes. For a reliable system, the ability to explain or describe its decisions is crucial; however, there tends to be a trade-off in that explainable systems cannot outperform non-explainable systems in terms of performance. In this paper, we propose incorporating the task of describing actions in language into the reinforcement learning of navigation as an auxiliary task. Existing studies have found it difficult to incorporate describing actions into reinforcement learning due to the absence of ground-truth data. We address this issue by leveraging knowledge distillation from pre-trained description generation models, such as vision-language models. We comprehensively evaluate our approach across various navigation tasks, demonstrating that it can describe actions while attaining high navigation performance. Furthermore, it achieves state-of-the-art performance in the particularly challenging multimodal navigation task of semantic audio-visual navigation.

1. Introduction

The development of robots that recognize their environment and autonomously navigate to a specified target has received particular attention in the last decade. Recently, multimodal navigation tasks have been proposed that can handle language [4, 53, 81] and can observe auditory information [9, 28]. Furthermore, various methods have been proposed to improve navigation performance, such as end-to-end reinforcement learning (RL) methods [4, 9, 25] and methods that utilize large language models (LLMs) [47, 73, 79]. As tasks become more complex and challenging, models become increasingly complex and black-boxed, making explainability crucial to addressing resulting lack of transparency. However, in general, prior research on explainability has been plagued by the trade-off between explainability

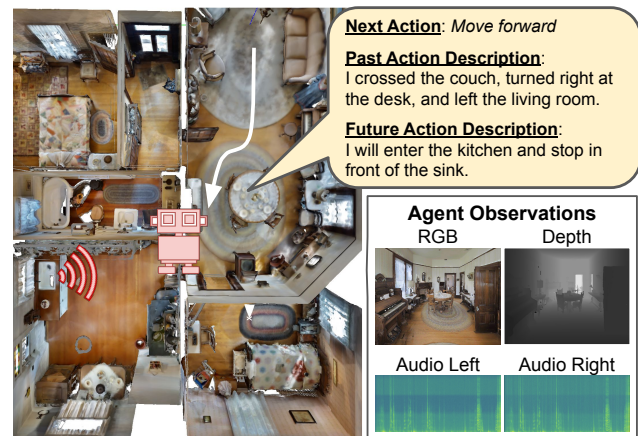


Figure 1. Overview of this study. The robot generates action descriptions of its past actions or future plans during navigation, such as moving toward a sounding object in the case of semantic audio-visual navigation, based on visual and auditory observations.

and the performance of the main task [21, 26, 57]. Using imitation learning (IL), there have been attempts to predict the instruction sentences given to the model [32, 80], in the context of instruction following tasks such as vision-and-language navigation [4]. However, the lack of ground-truth sentences in RL makes it difficult to extend these IL methods to RL, and this fact has limited the range of applications.

We propose a method to describe the system’s actions in natural language, aiming for an explainable navigation system. We define multiple types of action descriptions and focus on the affinity between this action description generation learning and navigation learning. Surprisingly, by formulating the action description as an auxiliary task through knowledge distillation from pre-trained models, such as large vision-language models (VLMs), to RL models, we developed a method that enables RL agents to describe actions without compromising navigation performance.

Figure 1 represents an overview of this research. During navigation, the RL agent also needs to describe what it has done in the past or what it should do in the future while making action decisions based on observations. We

demonstrate that the proposed method is applicable not only to a specific task (e.g., instruction following) but also to a wide range of navigation tasks, and it achieves state-of-the-art performance in a particularly challenging task, semantic audio-visual navigation (SAVNav) [11]. We also show that the proposed method makes the RL model descriptive and allows us to analyze failure cases. The main contributions of this paper are summarized as follows.

- Focusing on the compatibility between action description and navigation, we propose an RL method to learn both descriptive and high-performing navigation models.
- By leveraging pre-trained VLMs, our method removes the reliance on human-created data.
- A comprehensive experiment of various action descriptions and auxiliary task methods is conducted to analyze the proposed method.
- Through comparisons across various navigation tasks and existing methods, we demonstrate the consistent effectiveness of the proposed approach, achieving performance surpassing the state-of-the-art in the particularly challenging SAVNav task.

2. Related Work

Explainable Reinforcement Learning (XRL) is the framework where a reinforcement learning agent communicates its situation and reasons the decisions in a way humans can understand. Traditional approaches often involve transforming tasks into hierarchical structures by dividing them into smaller and more understandable subtasks for humans [7, 60], representing networks as tree structures [17, 43, 56], visualizing where the agent focuses on [30, 37, 41, 49, 61], or generating explanations in natural language [22, 23]. Recently, methods utilizing large language models (LLMs) to generate explanations related to agents' actions and observations have also been proposed [29, 66, 74]. Also, Stein *et al.* [62] addressed explainability in navigation by generating explanatory sentences in a rule-based method in the form of fill-in-the-blanks. Conventional methods have focused solely on explainability or interpretability, leading to a trade-off with the performance of the main task [21, 26, 57]. Similarly to previous studies [22, 23], we adopt a method that describes action decisions in natural language. However, our approach differs from theirs in that it integrates the action description generation module with the policy, which not only enhances the quality of the generated action description but also improves the performance of the navigation task.

Vision and Language Navigation (VLN) is a task of navigation based on first-person visual observations to follow instructions in a given natural language [4]. Previous research include splitting long instructions into shorter segments [31, 82], effectively integrating historical and multimodal information [13, 35, 40], optimizing the use of the

topological map [1, 2, 14], and enhancing performance by augmenting trajectory-instruction paired data [27, 38, 63, 68, 69]. Recently, methods for fine-tuning large pre-trained models [42, 50, 77] or utilizing them in a zero-shot manner [59, 78, 79] have also been proposed. The above approaches primarily use instructions as only inputs. Similar to our work, Zhu *et al.* [80] and Hejna *et al.* [32] use given instructions as outputs as well, treating instruction prediction as an auxiliary task. The most significant difference from our work is that they use only imitation learning, which has limitations such as a restricted exploration space, use of expensive demonstration data and dependency on human demonstration capabilities. It is not possible to simply transfer their method to RL because the ground-truth pair data for trajectory and instruction is not available in RL.

Audio Visual Navigation is a task that involves navigation to a sound location by observing auditory information in addition to visual information. Various RL-based methods have been proposed to tackle this task [9–12, 18, 48, 64, 67, 75, 76]. While these methods do not incorporate language, Paul *et al.* [51] and Liu *et al.* [44] introduced language into audio-visual navigation as instructions to reach the goal. More recently, methods leveraging LLMs have been proposed to more effectively utilize language in solving this task [36, 73]. Notably, Yang *et al.* [73] achieved higher success rates compared to learning-based methods. However, in terms of the SPL metric, which is considered crucial for navigation evaluation [3], their performance fell short of that of simple learning-based baselines.

3. Proposed Method

We propose descriptive reinforcement learning (DescRL), a method that enables to describe its own actions and improves navigation performance by treating action description as an auxiliary task. Here, we deal with three types of action description, past action description (P-AD), future action description (F-AD) and past-future action description (PF-AD) (Fig. 1). P-AD provides a verbalization of what the agent has done in the past. This helps to identify recognition errors in objects or spaces, or to verify that past information important for navigation is accurately remembered. Learning to predict P-AD will contribute to recognizing objects, spaces, etc., attention to important information, and acquiring common knowledge such as “the couch and the desk are in the living room.” F-AD provides a verbalization of what the agent should do in the future. This helps to verify whether the agent knows exactly what actions to take in the future, or whether the agent knows exactly but is unable to link them to actual actions. Learning to predict F-AD will contribute to the improvement of planning abilities and the acquisition of common sense, such as “the sink is in the kitchen.” Also, PF-AD is a combination of P-AD and F-AD. This provides a verbalization what the agent has

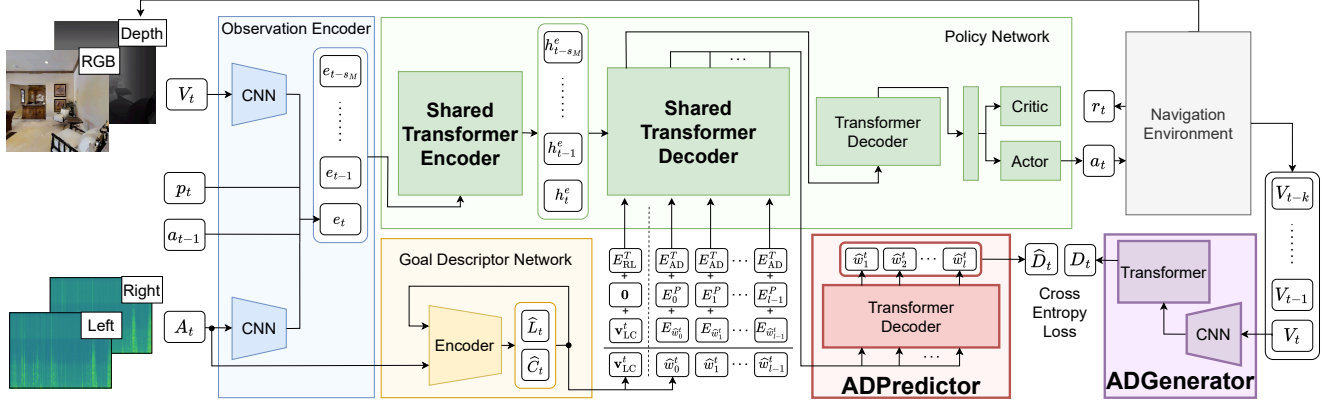


Figure 2. Overview of DescRL applied to SAVi [9], which is a method for semantic audio-visual navigation. The RL agent receives visual observation V_t , auditory observation A_t , posture p_t , and previous action a_{t-1} at time t and outputs the next action a_t and the action description \hat{D}_t . E^T and E^P represent task embedding and positional encoding, respectively. Here, ADPredictor is trained as an auxiliary task to predict the output D_t of the pre-trained ADGenerator.

done in the past and what the agent should do in the future.

DescRL is especially important for a complex task such as Semantic Audio-visual Navigation (SAVNav) [11]. In this task, the goal location must be identified even when auditory information is not observable because the sound may stop in the middle of the episode. In addition, if the sound stops in the middle of the episode, RL for navigation becomes difficult because the current observation is no longer enough to provide clues about the goal and cannot infer the reward. From this point of view, it is important to use auxiliary tasks for continuous learning. Experimentally, we show that DescRL is a particularly good auxiliary task compared to other auxiliary tasks.

Our proposed method is divided into two major phases. In phase 1, the ADGenerator, a module that verbalize navigation observations (bottom right of Fig. 2), is pre-trained to generate action descriptions from navigation observations. In phase 2, the navigation is trained, while the action description prediction is treated as an auxiliary task.

3.1. Phase 1: Pre-train ADGenerator

We first pre-train the ADGenerator, which takes the sequence of visual information observed by the agent and translates it into language. This is similar to a video captioning model or the speaker-model [27].

Structure in Object-goal Navigation (ObjNav) and SAV-Nav: The ADGenerator receives a one-hot vector sequence of the agent’s actions $\mathbf{a}_1, \dots, \mathbf{a}_T \in \{0, 1\}^4$ and the observed visual information sequence $V_0, \dots, V_T \in \mathbb{R}^{128 \times 128 \times 7}$. Here, in addition to RGBD images that can be observed by the RL agent, the ADGenerator also uses semantic images. The ADGenerator passes the visual information sequence through a convolutional neural network $f_V^g : \mathbb{R}^{128 \times 128 \times 7} \rightarrow \mathbb{R}^{512}$ to obtain the visual feature sequence $f_V^g(V_0), \dots, f_V^g(V_{T-1})$. Then, this visual feature sequence and the action sequence are concatenated to obtain

$(f_V^g(V_0), \mathbf{a}_1), \dots, (f_V^g(V_{T-1}), \mathbf{a}_T)$. Finally, by inputting this concatenated feature sequence to the transformer [65] encoder, the word sequence w_1, \dots, w_l is output from the transformer decoder. This word sequence is the verbalization of the visual information sequence V_0, \dots, V_T .

Structure in VLN: In VLN, the agent navigates on a graph. The ADGenerator receives the observation information v_0, \dots, v_K for each observed node and their neighbor information ϵ_t , which indicates whether each pair of nodes is adjacent. These are input to an encoder, which has the same structure as the Coarse-scale Cross-modal Encoder except for the Cross-Attention in DUET [14]. Then, the visual information features $\hat{v}_0, \dots, \hat{v}_K$ are obtained. Finally, by inputting these information into the transformer decoder, the sentence w_1, \dots, w_l is obtained.

Training details: The ADGenerator was trained using the R2R dataset [4, 40], which is a human-created dataset for vision and language navigation. The i -th data contains the navigation language instructions $w_1^i, \dots, w_{l_i}^i$, a sequence of visual information $V_0^i, \dots, V_{T_i}^i$ and an action sequence $\mathbf{a}_1^i, \dots, \mathbf{a}_{T_i}^i$ that can be observed by navigating along the given language instruction. So, the ADGenerator is trained to verbalize how the agent navigated from the input observations. We used 10,819 pieces of data for training and 1,839 pieces of data for evaluation. The training was performed using the teacher-forcing method [71]. The cross entropy loss was used as the loss function. As the word embedding, GloVe [52] was used for ObjNav and SAVNav, and BERT [20] tokenizer was used for VLN. See the supplementary materials for qualitative evaluation.

3.2. Phase 2: Action Description Prediction for Reinforcement Learning

After pre-training the ADGenerator, the RL model learns navigation while treating action description prediction as an auxiliary task. The module is called ADPredictor. The RL

model learns to output appropriate actions from the policy network as before but also learns to output action descriptions from the ADPredictor. Figure 2 shows the overview of the application of DescRL to SAVi [11], which is a method for SAVNav. SAVi has a CNN-based observation encoder, a transformer-based policy network, and a goal descriptor network that predicts goal location L_g and category C_g . Concatenating them, the input to the policy network $v_{LC}^t = (\hat{L}_t, \hat{C}_t)$ is calculated. ADPredictor consists of a transformer decoder. By inputting the predicted goal location \hat{L}_g and category \hat{C}_g of the goal as the beginning of sentence (\hat{w}_0^t in Fig. 2), an action description is predicted considering \hat{L}_g, \hat{C}_g . Here, the transformer encoder and decoder are shared by ADPredictor and the policy. This allows the shared encoder and decoder to be trained on two tasks so that it can encode observation information more effectively. Also, it is expected to be able to predict more faithful [45] action descriptions for the model since they share most of the weights. To enable decoder sharing, task embeddings E_{RL}^T, E_{AD}^T are applied to the input for policy and ADGenerator, respectively. This acts like the positional encoding commonly used in transformer. In addition, visual observations for the past $k + 1$ steps are input to the pre-trained ADGenerator. As a result, the ADGenerator outputs an action description of what the agent has done in the past. ADPredictor is trained with cross-entropy loss \mathcal{L}^{CE} to be able to predict this ADGenerator output. Therefore, if the loss function for RL is \mathcal{L}^{RL} , this model is trained to minimize $\mathcal{L}^{RL} + \lambda \mathcal{L}^{CE}$, where λ is a coefficient indicating how much to account for the loss of action description prediction. For other architectures of other tasks, the ADPredictor and ADGenerator were applied in the same way (see the supplementary materials for more details).

The training of this RL model is divided into two steps.

Step 1: Pre-training of ADPredictor. Here, this model does not learn navigation but only action description prediction to learn the observation encoder, the shared encoder and decoder, and ADPredictor.

Step 2: By learning navigation and action description prediction at the same time, both the entire policy network and ADPredictor are learned. Here, the ADPredictor is always trained in a teacher-forcing manner [71].

For training in Step 1, R2R can be utilized in VLN, but not in SAVNav and ObjNav. Therefore, we created a new dataset with the following operations. First, the observation O_0, O_1, \dots, O_T is obtained by taking the shortest path from a random start to a random goal, and action sequences $\mathbf{a}_1, \dots, \mathbf{a}_T$ are collected. Then, these information are input to the ADGenerator learned in phase 1 to obtain the ground truth action description D . ADPredictor is trained using the dataset consisting of approx. 100k data for ObjNav and 500k data for SAVNav created in advance using this method. The ADGenerator is used only

during training and not during testing so that future observations can be input. That is, instead of V_{t-k}, \dots, V_t , the visual observations V_t, \dots, V_{t+k} obtained by following the shortest path to the goal can be input to ADGenerator. Here, the output D_t of the ADGenerator represents what to do in the future. In this case, the ADPredictor must predict what to do in the future based on past observations and the predicted location and category of the goal. In the following, we refer to DescRL that allows past/future/past-and-future observations to be input to the ADGenerator as Past-DescRL/Future-DescRL/Past-and-Future-DescRL (P-DescRL/F-DescRL/PF-DescRL), respectively. Future observations are only used at train time through ADGenerator. At test-time, the ADPredictor only has access to past observations, ensuring fairness with prior studies

Training details in ObjNav: We used 8 GPUs and assigned 10 processes to each GPU, for a total of 80 processes for training. The RL algorithm used for training was DD-PPO [70]. The number of parameter updates was set to 3,000. The number of transformer encoder layers was set to 2, and the number of transformer shared decoder layers was set to 2. The number of unshared decoder layers was set to 1 in both the policy and ADPredictor. The coefficient of DescRL loss was set to $\lambda = 0.1$, and the input length to ADGenerator was set to $k + 1 = 20$.

Training details in VLN: We used 1 GPU and assigned 1 process to the GPU for training. The imitation learning (IL) algorithm DAGger [55] was used for training. Our method is applicable not only to RL but also to IL. The number of transformer shared decoder layers was set to 0 because baseline methods did not have any decoder. The number of unshared decoder layers was set to 3 in ADPredictor. The coefficient of DescRL loss was set to $\lambda = 0.1$, and all visual observation history is input to ADGenerator.

Training details in SAVNav: We used 4 GPUs and assigned 8 processes to each GPU, for a total of 36 processes for training. Other settings related to DescRL are the same as for ObjNav. More details can be found in the supplementary materials and the code to be published.

3.3. Vision language models for ADGenerator

In the above, we proposed that the ADGenerator is trained from scratch on the human-created VLN dataset R2R [4, 40]. This means that the proposed method still relies on human-created data, which undermines its applicability. Here, we also propose an alternative approach based on knowledge distillation techniques [33] by focusing on vision language models (VLMs) that have already acquired common sense from large-scale web data.

We used VideoLLaMA2 [15] and QWen2.5-VL [6] as VLM. It receives an input consisting of RGB images $(V_1, \dots, V_v) \in \mathbb{R}^{l^v \times 336 \times 336 \times 3}$ and a prompt (see the supplementary materials for details) The output is treated as the ground-truth data for ADPredictor’s output. Knowledge

distillation was conducted in the form of soft targets [33]. Zhou *et al.* [78] pointed out that navigation methods that use foundation models in zero-shot cannot beat learning-based methods, while methods that fine-tune foundation models have the problem of reduced sentence generation ability due to catastrophic forgetting. Our method further improves the performance of learning-based methods and does not cause catastrophic forgetting. Furthermore, our method is lightweight compared to foundation models, making it capable of running in real time.

Additionally, we experimented with leveraging VideoL-LaMA2 not only in a zero-shot setting but also by fine-tuning it on the R2R dataset [4, 40] using QLoRA [19]. The number of epochs was set to 1, the learning rate to 2.0×10^{-5} , LoRA r to 128, and LoRA α to 256. See the supplementary materials for more details.

4. Experiments

4.1. Implementation Details

Object-goal Navigation: Habitat Simulator [58] and Matterport3D [8] scene dataset were used to train RL agents. The dataset consists of large floors with an average of 512 m². We divide the 67 scenes into 56/4/7 splits for train/val/test, respectively. If \mathbb{I}_{goal} represents whether or not the agent has reached the goal, d_t represents the geodesic distance from the agent to the goal at time t , and r_{penalty} represents the time penalty, then the reward at time t can be expressed as $r_t = \alpha \mathbb{I}_{\text{goal}} + (d_t - d_{t-1}) + r_{\text{penalty}}$. Here, the coefficient α and the time penalty are set to 2.5, -0.001 . We used three commonly used metrics: Success Rate (SR), Success rate weighted by Path Length (SPL), Distance To Goal (DTG). SR represents the rate of the agent reaching goal. SPL is the standard navigation metric [3], which is the success rate weighted by the ratio of the shortest path to the actual agent path length. In other words, success with a route that is close to the shortest path results in a higher SPL. DTG is the distance from the agent location at the end of the episode to the goal.

Vision-and-language Navigation: Matterport3DSimulator [4] was used to train IL agents. The entire setup is exactly the same as in the previous study [14, 69]. The same evaluation metrics are used as in the previous study [69]; Navigation Error (NE), SR, and SPL. Here, NE is same as DTG.

Semantic Audio-visual Navigation: Soundspaces [9] simulator and Matterport3D [8] scene dataset were used to train RL agents. We divide the 102 scenes into 73/11/18 splits for train/val/test, respectively. The setup is exactly the same as in the previous study [11]. Using the indicator function $\mathbb{I}[\cdot]$, the reward at time t can be expressed as $r_t = \alpha \mathbb{I}_{\text{goal}} + \mathbb{I}[d_t > d_{t-1}] + r_{\text{penalty}}$. Here, α and r_{penalty} are set to 10, -0.01 , respectively. The same evaluation metrics are used as in the previous study [9]. That is, SR, SPL, Success rate weighted by Number of Actions (SNA), DTG,

Table 1. Comparison on object-goal navigation. PT indicates whether or not the ADPredictor pre-training (step 1 of phase 2) has been done. N_{SD} indicates the number of sharing decoders.

	Method	PT	N_{SD}	SR \uparrow	SPL \uparrow	DTG \downarrow
1)	GRU-based [58]	-	-	7.4	4.2	6.49
2)	SMT [25]	-	-	17.9	7.7	6.72
3)	SMT w/ Past-DescRL	\times	0	12.8	5.7	6.95
4)	SMT w/ Past-DescRL	\checkmark	0	26.7	9.7	5.91
5)	SMT w/ Past-DescRL	\checkmark	2	22.8	8.0	5.91

Table 2. Comparison on vision-and-language navigation.

Method	Val Seen			Val Unseen			Test Unseen		
	NE \downarrow	SR \uparrow	SPL \uparrow	NE \downarrow	SR \uparrow	SPL \uparrow	NE \downarrow	SR \uparrow	SPL \uparrow
Human	-	-	-	-	-	-	1.61	86	76
Random	9.49	16.27	14.91	9.22	16.01	14.01	-	-	-
Seq2Seq [4]	5.87	37.45	32.36	8.01	21.24	18.00	-	-	-
DUET [14]	2.26	79.73	74.78	3.21	71.65	60.44	3.63	69.76	59.39
w/ Future-DescRL	2.43	80.02	74.03	3.20	71.73	60.04	-	-	-
w/ Past-DescRL	2.12	81.00	76.27	3.09	72.33	61.37	3.56	69.88	59.40
ScaleVLN [69]	1.95	82.57	77.09	2.40	78.63	69.15	2.61	77.14	67.34
w/ Future-DescRL	2.02	81.29	75.78	2.44	78.46	69.48	-	-	-
w/ Past-DescRL	1.88	82.66	76.46	2.37	78.84	68.96	2.64	77.47	66.96

and Success When Silent (SWS). SNA is the success rate weighted by the ratio of the minimum number of actions and the number of agent actions. SWS is the success rate when the sound stops. In this study, the average value over 1,000 tests was obtained.

4.2. Navigation Performance

4.2.1 Object-goal Navigation

Two baselines were used. **1) GRU-based [58]** is the simplest method for this task. It is based on GRU [16] and is trained with end-to-end RL. **2) SMT [25]** has a transformer-based structure and learns by RL. By keeping visual information in a memory, this considers history.

We compared the proposed method Past-DescRL with the baseline methods. The proposed method was applied to SMT [25]. As shown in Table 1, our method improved baseline performance on all metrics.

4.2.2 Vision-and-language Navigation

Four baselines were used. **1) Random** turns to a randomly selected heading and then moves forward. In total, it takes five actions. **2) Seq2Seq [4]** is the simplest model with an encoder-decoder structure. It outputs actions by inputting an instruction into the encoder and visual observations into the decoder. **3) DUET [14]** has a dual-scale graph transformer for simultaneous long-term action planning and detailed cross-modal understanding. It builds a topological map in real-time to enable efficient exploration in the global action space. **4) ScaleVLN [69]** is a data augmentation method using HM3D [54] and Gibson [72], scene datasets that are different from Matterport3D used in R2R.

This achieved state-of-the-art by combining with DUET.

We compared the proposed method Past-DescRL and Future-DescRL with the baseline methods. The proposed method was applied to DUET [14] and ScaleVLN [69]. As shown in Table 2, Past-DescRL slightly improves performance on all metrics for DUET, and the performance improvement is even smaller for ScaleVLN. Since the proposed method is an auxiliary method, its contribution to the main task may be considered small if the task is simple or if the method already performs well. While it is true that the SPL of ScaleVLN is not high enough, we believe the difficulty of navigation tasks cannot be fully captured by SPL alone. It also involves the complexity of action decision-making from the robot’s perspective. In this regard, SAVNav presents a particular challenge, as the sound stops midway, making it difficult to determine appropriate actions in the latter part of the task. In contrast, tasks such as ObjNav and VLN consistently provide a target object or a language instruction, which facilitates easier decision-making. Consequently, we observe a significant performance improvement in SAVNav, where both the baseline performance and decision-making ease are lower than the other tasks. On the other hand, it is noteworthy that the proposed method does not reduce the performance of the main task, even though it adds the ability to generate action descriptions.

4.2.3 Semantic Audio-visual Navigation

Five baselines were used. **1) Random** randomly selects an action from $\{MoveForward, TurnLeft, TurnRight\}$ and *Stop* if it reaches within a 1 m radius of the goal. **2) AV-Nav [9]** is the simplest method first proposed for audio-visual navigation. It is based on GRU and is trained with end-to-end RL. **3) AV-WaN [10]** learns to generate waypoints using map information. This one is also based on GRU and is learned by end-to-end RL. **4) SAVi [11]** is the first method for semantic audio-visual navigation. It has a transformer-based structure and learns by RL. It also has a module called Goal Descriptor Network, which predicts goal position and goal category from auditory observations. **5) KSAVEN [64]** is a knowledge-driven method that uses prior knowledge by utilizing a pre-created knowledge graph. It has a transformer-based structure and learns by RL. The knowledge graph is processed by using a graph convolution network [39]. This achieved state-of-the-art performance. The original paper [64] states that all sounds (including unheard) are used to pre-train the audio network, but in this study, only heard sounds were used for fairness with other baselines and making the setting realistic.

We compared the proposed methods P-DescRL, F-DescRL and PF-DescRL with the baseline methods. The proposed method was applied to SAVi [11] and KSAVEN [64]. As shown in Table 3, our method achieved state-of-the-art on all evaluation metrics. We also found that learn-

Table 3. Comparison of state-of-the-art methods and DescRL. In the Heard/Unheard settings, we tested using sounds heard/not heard during training, respectively.

Method	Heard					Unheard				
	SR↑	SPL↑	SNA↑	DTG↓	SWS↑	SR↑	SPL↑	SNA↑	DTG↓	SWS↑
Random	6.4	1.8	0.7	16.5	6.2	6.4	1.8	0.7	16.5	6.2
AV-Nav [9]	19.3	15.9	15.0	12.6	5.6	15.7	12.8	11.9	12.6	4.7
AV-WaN [10]	15.9	12.3	11.6	11.0	6.1	12.5	9.9	9.2	11.2	5.1
SAVi [11]	31.6	28.5	24.6	11.8	12.5	24.7	22.4	18.9	11.8	10.2
w/ F-DescRL	36.4	30.8	26.6	8.4	17.7	22.5	19.1	16.8	9.0	9.8
w/ PF-DescRL	32.6	27.3	22.2	9.7	17.0	29.4	24.4	19.5	10.2	15.2
w/ P-DescRL	37.4	32.4	28.0	8.4	19.1	31.4	26.9	22.5	8.7	15.1
KSAVEN [64]	25.1	18.1	13.5	10.3	15.8	21.1	14.9	11.5	11.2	14.2
w/ F-DescRL	27.4	20.4	17.2	12.2	14.4	20.6	14.9	12.6	12.8	10.6
w/ P-DescRL	21.4	15.4	11.2	12.4	13.8	21.2	15.3	10.9	12.0	14.5

ing to generate P-AD has a better impact on navigation than learning to generate F-AD. We believe this is because F-AD prediction is a difficult task in itself and therefore unsuitable as an auxiliary task. We also find that F-DescRL performs worse than the baseline in unheard setting. F-DescRL performs better when it is easy to infer what to do in the future, but not when it is hard to infer what to do in the future. Thus, in the unheard setting, where it is difficult to predict goal categories from sound and hard to infer what to do in the future, performance is expected to be worse than in the baseline. Conversely, F-DescRL outperforms PF-DescRL in heard setting because goal categories are easier to infer and it is easier to infer what to do in the future.

Qualitative results Figure 3 shows the results of the qualitative evaluation for navigation of SAVi [11] w/ P-DescRL. The evaluation was based on unseen scenes and unheard sounds during training. This result shows that the P-DescRL reduces wasteful actions and enables more accurate navigation. We believe this is because the action description prediction task allows the agent to learn more important knowledge for navigation. The bottom right of Fig. 3 shows a case of P-DescRL failure. In P-DescRL, there were often failure cases where the agent went close but could not stop at the exact position. We believe this is due to the domain gap between VLN and SAVNav. R2R used to train ADGenerator is the dataset for VLN. Here, the distance to the goal considered successful in VLN is 3 m, whereas the distance considered successful in SAVNav is 1 m. In subsequent experiments, SAVNav is used unless otherwise noted.

4.2.4 Comparison to other auxiliary tasks

We investigated whether action description prediction is a particularly effective auxiliary task compared to other alternatives. For comparison, we used a supervised learning approach to learn to predict the next action, the progress [46], the next image observation, the next auditory observation, the goal location, and the goal category. The results are shown in Table 4. Since the proposed method outperforms

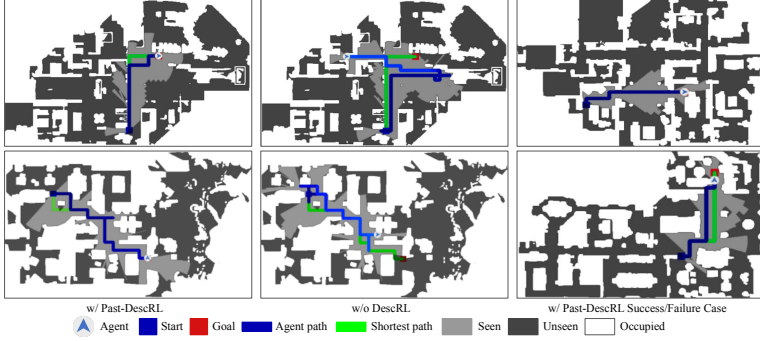


Figure 3. Navigation trajectories. Comparing the left and middle figures shows that DescRL reduces wasteful actions. The bottom right figure shows a failure case where the agent got close to the goal but failed to stop at the exact position.

Table 4. Comparison with other auxiliary tasks.

Method	Heard				Unheard					
	SR↑	SPL↑	SNA↑	DTG↓	SWS↑	SR↑	SPL↑	SNA↑	DTG↓	SWS↑
SAVi [11]	31.6	28.5	24.6	11.8	12.5	24.7	22.4	18.9	11.8	10.2
w/ Next Action	33.2	30.6	27.3	10.7	12.7	23.1	20.8	17.8	11.1	9.3
w/ Progress [46]	35.0	31.6	28.3	10.2	15.7	24.2	21.1	18.3	11.0	9.4
w/ Next Frame	35.4	31.6	27.0	11.0	15.8	24.7	21.8	18.4	11.5	9.9
w/ Next Spectrogram	34.7	31.3	27.3	11.4	13.4	23.9	21.0	18.1	12.0	8.8
w/ Goal Location	34.5	30.2	25.8	10.3	15.6	25.7	22.3	18.5	10.7	11.2
w/ Goal Category	35.4	31.9	27.9	10.2	15.2	25.3	22.4	19.4	10.8	10.5
w/ F-DescRL	36.4	30.8	26.6	8.4	17.7	22.5	19.1	16.8	9.0	9.8
w/ PF-DescRL	32.6	27.3	22.2	9.7	17.0	29.4	24.4	19.5	10.2	15.2
w/ P-DescRL	37.4	32.4	28.0	8.4	19.1	31.4	26.9	22.5	8.7	15.1

the others in most of the metrics, it is clear that the proposed method is a particularly good auxiliary task. A semantic understanding of the environment is important for navigation, and we believe that the proposed method promotes this. The proposed method is highly beneficial in terms of not only improving navigation performance but also outputting natural language that can be interpreted by humans.

4.2.5 VLM as ADGenerator

The results are shown in Table 5. The best performance was found when using the ADGenerator, which has a simple structure learned from scratch using R2R. However, even in the case without any human-created data (row 3), we found that DescRL can improve the performance. This indicates that our method is able to implicitly extract the knowledge necessary for navigation from VLMs without any human-created data. We also found that the performance was lower when we used the VLM fine-tuned by R2R dataset. This may be due to the VLM being overfitted to the dataset. Furthermore, when comparing row 3 and row 5, using a stronger foundation model (Qwen2.5-VL) does not necessarily lead to better performance.

4.3. Analysis of Failures by Action Description

Here, we focus on the action descriptions generated by the ADPredictor in the failed episode and analyze why it failed.

Agent: go down the deck and then turn slight right and go past the chairs and wait near the couch.



Figure 4. Qualitative evaluation of action descriptions. The agent generates the above action descriptions when it observes the RGBD image.

Table 5. Comparison of using a model trained from scratch and using a pre-trained VLM as the ADGenerator (ADGen). CNN-TF, VL2, and QVL indicate CNN+Transformer, VideoLLaMA2, and Qwen2.5-VL respectively. FT indicates whether the model was fine-tuned (or trained from scratch) by the R2R dataset.

	Heard				Unheard							
	ADGen	FT	SR↑	SPL↑	SNA↑	DTG↓	SWS↑	SR↑	SPL↑	SNA↑	DTG↓	SWS↑
1) None	-		31.6	28.5	24.6	11.8	12.5	24.7	22.4	18.9	11.8	10.2
2) CNN+TF	✓	37.4	32.4	28.0	8.4	19.1	31.4	26.9	22.5	8.7	15.1	
3) VL2	×	33.7	29.8	24.4	10.6	16.0	28.9	24.3	19.5	10.6	14.2	
4) VL2	✓	28.9	25.6	21.6	11.6	11.6	23.3	20.3	16.7	11.9	10.2	
5) QVL	×	33.4	28.6	25.6	8.9	15.2	28.2	23.8	20.5	9.3	11.5	



Figure 5. Qualitative evaluation of generating action descriptions performance of ScaleVLN [69] w/ Past-DescRL on VLN. The agent observes the panoramic image at each step and must follow the given instructions. The red arrows represent the actions selected by the agent, and the action descriptions shown above each panorama image are ones generated by the agent. Areas marked in green are action descriptions that can be judged to be correct.

It generates action description based on its own output, the probability of each word $\hat{p}^S(w_l = w|w_1, \dots, w_{l-1})$, ($w \in \text{Vocab}$), using greedy search in VLN and using top-k sampling [24] and top-p sampling [34] where the temperature $\tau = 2.0$ and $k = 10$, $p = 0.95$ in SAVNav.

Figure 4 shows the RGBD image observed by the agent

and the action description during the failure episode in the bottom right in Fig. 3. The agent outputs “wait near the couch” even though it is still far from the goal chair framed in red. It can be analyzed that the agent stopped early because it judged that it was close enough to the chair, even though it was still far.

Figure 5 shows a result of the qualitative evaluation of action descriptions of ScaleVLN [69] w/ Past-DescRL on VLN. Here, it is evaluated with Val Unseen dataset. In the episode in Fig. 5, the agent stopped at the 7th step, but actually had to stop at 5th step. In other words, as the instruction says, the agent needed to stop on the floor circle enclosed by the red square, but was unable to do so. When we look at the action descriptions predicted by the agent, we can see that it did not pay attention to the circle on the floor at all. Therefore, it can be analyzed that the agent could not stop at the correct position because it could not pay attention to the circle on the floor.

Quantitative evaluation and more detailed qualitative evaluation can be found in the supplementary material.

4.4. Ablation Study

ADPredictor Pre-training: Comparing rows 1 and 2, rows 4 and 6 in Table 6, and rows 3 and 4 in Table 1, we can see that the performance is always better with pre-training (step 1 of phase 2). Therefore, we can see that this is a very important factor. This is because the learning unrelated to navigation, such as grammar, is done in advance.

Task Embedding: Comparing rows 5 and 6 of Table 6, we can see that the performance is higher with task embeddings for most of the metrics. Thus, we can see that this is also an important factor. This is important in determining whether the agent is now trying to decide on an action or predict an action description.

The number of sharing decoders: We investigate how many layers of shared decoders would improve SAVNav performance the most when an agent has three layers of decoders. Comparing rows 2, 3, 6, and 7 of Table 6, we found that sharing two out of three layers improves SAVNav performance for many of the metrics. If no decoders are shared at all, the benefit to navigation from action description prediction is reduced. On the other hand, if all decoders are shared, the latent space for decision-making and the latent space for predicting action descriptions must be equivalent. However, this is contrary to the fact. Therefore, conversely, the case where all decoders are shared would also not have resulted in higher accuracy. This result indicates that a trade-off still exists between navigation and faithfulness. Also, in ObjNav, the performance was higher when the number of shared decoders was smaller (row 4 and row 5 in Table 1). This suggests that SAVNav is a particularly difficult task, and therefore it is especially important to supplement it with other tasks.

Table 6. Results of the ablation experiment. PT indicates whether the ADPredictor is pre-trained. TE indicates whether task embedding is used. N_{SD} indicates the number of sharing decoders.

	PT	TE	N_{SD}	Heard					Unheard				
				SR \uparrow	SPL \uparrow	SNA \uparrow	DTG \downarrow	SWS \uparrow	SR \uparrow	SPL \uparrow	SNA \uparrow	DTG \downarrow	SWS \uparrow
1)	\times	-	0	33.9	30.7	26.8	10.4	13.6	25.6	22.2	19.2	10.9	10.4
2)	\checkmark	-	0	37.6	32.6	28.0	7.8	18.6	26.6	23.7	20.3	8.7	11.2
3)	\checkmark	\checkmark	1	38.7	33.3	27.2	8.0	21.5	30.5	26.1	20.9	8.7	16.2
4)	\times	\checkmark	2	34.1	31.3	27.6	10.3	13.3	25.1	22.3	19.4	11.0	10.4
5)	\checkmark	\times	2	37.4	31.7	25.0	8.3	21.6	30.0	24.9	20.0	8.7	16.9
6)	\checkmark	\checkmark	2	37.4	32.4	28.0	8.4	19.1	31.4	26.9	22.5	8.7	15.1
7)	\checkmark	\checkmark	3	30.2	26.6	19.8	9.9	15.9	26.5	23.0	17.1	10.3	14.1

4.5. Limitations

The limitations of our current proposed method are as follows. 1) Harm caused by biased common sense in the dataset. For example, due to the frequent occurrence of stairs in the hallways in the dataset, the word “stairs” appeared even if there are no actual stairs when walking down the hallway. 2) Stopping in front of an object in the true category but in different instances. Due to the lack of qualifiers about the object in the description, the category can be identified but the specific appearance cannot be learned. In our proposed method using VLMs, we can expect improvement by adjusting the prompt to generate action descriptions rich in modifiers. 3) Adaptation beyond navigation. We are interested in whether the proposed method can also be applied to other RL tasks or to Embodied Question Answering [5]. 4) The proposed method may not work if the robot differs from those assumed when creating the training data for generating action descriptions. For example, it is questionable whether the ADGenerator trained on the R2R dataset, which was created assuming a roughly human-sized robot, can be applied to learning a navigation model for a smaller robot. The view changes with the robot’s size, so the correct action description should also adjust accordingly.

5. Conclusion

We proposed DescRL, an approach that enables agents to describe their actions while integrating both IL and RL methods. Our experiments confirmed that action description prediction tends to enhance navigation performance more effectively than other auxiliary tasks. Also, by leveraging knowledge distillation from a VLM, our experiments suggested that DescRL can be trained without reliance on expensive human-created data, making it a more adaptable method. We demonstrated its consistent effectiveness for various navigation tasks in improving navigation performance while making the navigation descriptive. Furthermore, our method achieved state-of-the-art performance on SAVNav, a highly challenging navigation task, and suggested the possibility of analyzing failure cases using the predicted action descriptions.

Acknowledgments: This work was supported by JST FOREST Program, Grant Number JPMJFR206H.

References

- [1] Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. Bevbort: Multimodal map pre-training for language-guided navigation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2737–2748, 2023. 2
- [2] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [3] 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. 2, 5
- [4] 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 *CVPR*, 2018. 1, 2, 3, 4, 5
- [5] Xiaohan Zhang Pranav Putta Sriram Yenamandra Mikael Henaff Sneha Silwal Paul Mcvay Oleksandr Maksymets Sergio Arnaud Karmesh Yadav Qiyang Li Ben Newman Mohit Sharma Vincent Berges Shiqi Zhang Pulkit Agrawal Yonatan Bisk Dhruv Batra Mrinal Kalakrishnan Franziska Meier Chris Paxton Sasha Sax Aravind Rajeswaran Arjun Majumdar, Anurag Ajay. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *CVPR*, 2024. 8
- [6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv*, 2025. 4
- [7] Benjamin Beyret, Ali Shafti, and A. Aldo Faisal. Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 5014–5019. IEEE Press, 2019. 2
- [8] 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 *Proceedings of Conference on 3D Vision (3DV)*. IEEE, 2017. 5
- [9] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vincenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 17–36. Springer, 2020. 1, 2, 3, 5, 6
- [10] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020. 6
- [11] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *CVPR*, 2021. 2, 3, 4, 5, 6, 7
- [12] Jinyu Chen, Wenguan Wang, Si Liu, Hongsheng Li, and Yi Yang. Omnidirectional information gathering for knowledge transfer-based audio-visual navigation. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 10993–11003, 2023. 2
- [13] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 5834–5847, 2021. 2
- [14] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16537–16547, 2022. 2, 3, 5, 6
- [15] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv*, 2024. 4
- [16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NeurIPS Workshop on Deep Learning*, 2014. 5
- [17] Youri Coppens, Kyriakos Efthymiadis, Tom Lenaerts, and Ann Nowé. Distilling deep reinforcement learning policies in soft decision trees. In *Proceedings of International Joint Conference on Artificial Intelligence*, 2019. 2
- [18] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. See, hear, explore: Curiosity via audio-visual association. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 14961–14972, 2020. 2
- [19] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 5
- [20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. 3
- [21] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *Proceedings of International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, 2018. 1, 2
- [22] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society (AI/ES)*, pages 81–87, 2018. 2
- [23] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of International Conference on Intelligent User Interfaces (IUI)*, pages 263–274, 2019. 2
- [24] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual*

- Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, 2018. 7
- [25] Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 538–547, 2019. 1, 5
- [26] Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD Explorations*, 15(1):1–10, 2014. 1, 2
- [27] 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. In *NeurIPS*, 2018. 2, 3
- [28] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE, 2020. 1
- [29] Miguel Á. González-Santamarta, Laura Fernández-Becerra, David Sobrín-Hidalgo, Ángel Manuel Guerrero-Higueras, Irene González, and Francisco J. Rodríguez Lera. Using large language models for interpreting autonomous robots behaviors. In *Hybrid Artificial Intelligent Systems*, pages 533–544, Cham, 2023. Springer Nature Switzerland. 2
- [30] Lei He, Nabil Aouf, and Bifeng Song. Explainable deep reinforcement learning for uav autonomous path planning. *Aerospace Science and Technology*, 118:107052, 2021. 2
- [31] Zongtao He, Liuyi Wang, Shu Li, Qingqing Yan, Chengju Liu, and Qijun Chen. Mlanet: Multi-level attention network with sub-instruction for continuous vision-and-language navigation. *arXiv preprint arXiv:2303.01396*, 2023. 2
- [32] Joey Hejna, Pieter Abbeel, and Lerrel Pinto. Improving long-horizon imitation through instruction prediction. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 7857–7865, 2023. 1, 2
- [33] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4, 5
- [34] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020. 7
- [35] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653, 2021. 2
- [36] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Audio visual language maps for robot navigation. In *International Symposium on Experimental Robotics*, pages 105–117. Springer, 2023. 2
- [37] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on explainable artificial intelligence*, 2019. 2
- [38] Aishwarya Kamath, Peter Anderson, Su Wang, Jing Yu Koh, Alexander Ku, Austin Waters, Yinfei Yang, Jason Baldrige, and Zarana Parekh. A new path: Scaling vision-and-language navigation with synthetic instructions and imitation learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10813–10823, 2023. 2
- [39] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. 6
- [40] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 104–120. Springer, 2020. 2, 3, 4, 5
- [41] Edouard Leurent and Jean Pierre Mercat. Social attention for autonomous decision-making in dense traffic. *ArXiv*, abs/1911.12250, 2019. 2
- [42] 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. 2
- [43] Guiliang Liu, Oliver Schulte, Wang Zhu, and Qingcan Li. Toward interpretable deep reinforcement learning with linear model u-trees. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part II 18*, pages 414–429. Springer, 2019. 2
- [44] Xiulong Liu, Sudipta Paul, Moitreyia Chatterjee, and Anoop Cherian. Caven: An embodied conversational agent for efficient audio-visual navigation in noisy environments. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2023. 2
- [45] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723, 2024. 4
- [46] Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019. 6, 7
- [47] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems*, 35:32340–32352, 2022. 1
- [48] Sagnik Majumder and Shailesh Mani Pandey. Semantic audio-visual navigation through distractor silencing. 2
- [49] Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. Towards interpretable reinforcement learning using attention augmented agents. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2
- [50] Bowen Pan, Rameswar Panda, SouYoung Jin, Rogerio Feris, Aude Oliva, Phillip Isola, and Yoon Kim. Langnav: Language as a perceptual representation for navigation. In *Findings of the Association for Computational Linguistics*, pages 950–974, 2024. 2

- [51] Sudipta Paul, Amit Roy-Chowdhury, and Anoop Cherian. Avlen: Audio-visual-language embodied navigation in 3d environments. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 6236–6249, 2022. 2
- [52] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3
- [53] 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
- [54] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 5
- [55] 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. 4
- [56] Aaron M Roth, Nicholay Topin, Pooyan Jamshidi, and Manuela Veloso. Conservative q-improvement: Reinforcement learning for an interpretable decision-tree policy. *arXiv preprint arXiv:1907.01180*, 2019. 2
- [57] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019. 1, 2
- [58] 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 International Conference on Computer Vision (ICCV)*, pages 9339–9347, 2019. 5
- [59] Dhruv Shah, Błażej Osipiński, Sergey Levine, et al. Lmnav: Robotic navigation with large pre-trained models of language, vision, and action. In *Proceedings of Conference on Robot Learning (CoRL)*, pages 492–504. PMLR, 2023. 2
- [60] Tianmin Shu, Caiming Xiong, and Richard Socher. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2018. 2
- [61] Julian Skirzyński, Frederic Becker, and Falk Lieder. Automatic discovery of interpretable planning strategies. *Machine Learning*, 110:2641–2683, 2021. 2
- [62] Gregory Stein. Generating high-quality explanations for navigation in partially-revealed environments. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 17493–17506, 2021. 2
- [63] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2610–2621, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2
- [64] Gyan Tatiya, Jonathan Francis, Luca Bondi, Ingrid Navarro, Eric Nyberg, Jivko Sinapov, and Jean Oh. Knowledge-driven scene priors for semantic audio-visual embodied navigation. *arXiv preprint arXiv:2212.11345*, 2022. 2, 6
- [65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3
- [66] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv: Arxiv-2305.16291*, 2023. 2
- [67] Hongcheng Wang, Yuxuan Wang, Fangwei Zhong, Mingdong Wu, Jianwei Zhang, Yizhou Wang, and Hao Dong. Learning semantic-agnostic and spatial-aware representation for generalizable visual-audio navigation. *IEEE Robotics and Automation Letters*, 2023. 2
- [68] Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Izzeddin Gur, Natasha Jaques, Austin Waters, Jason Baldrige, and Peter Anderson. Less is more: Generating grounded navigation instructions from landmarks. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15407–15417, 2021. 2
- [69] 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. 2, 5, 6, 7, 8
- [70] 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 *Proceedings of International Conference on Learning Representations (ICLR)*, 2019. 4
- [71] Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989. 3, 4
- [72] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9068–9079, 2018. 5
- [73] Zeyuan Yang, Jiageng Liu, Peihao Chen, Anoop Cherian, Tim K Marks, Jonathan Le Roux, and Chuang Gan. Rila: Reflective and imaginative language agent for zero-shot semantic audio-visual navigation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16251–16261, 2024. 1, 2

- [74] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2023. 2
- [75] Yinfeng Yu, Lele Cao, Fuchun Sun, Xiaohong Liu, and Liejun Wang. Pay self-attention to audio-visual navigation. In *Proceedings of British Machine Vision Conference (BMVC)*. BMVA Press, 2022. 2
- [76] Yinfeng Yu, Wenbing Huang, Fuchun Sun, Changan Chen, Yikai Wang, and Xiaohong Liu. Sound adversarial audio-visual navigation. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022. 2
- [77] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13624–13634, 2024. 2
- [78] Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2024. 2, 5
- [79] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *AAAI*, 2024. 1, 2
- [80] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10012–10022, 2020. 1, 2
- [81] 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
- [82] Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, Eugene Ie, and Fei Sha. Babywalk: Going farther in vision-and-language navigation by taking baby steps. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2539–2556, 2020. 2