CAVEN: An Embodied Conversational Agent for Efficient Audio-Visual Navigation in Noisy Environments

Liu, Xiulong; Paul, Sudipta; Chatterjee, Moitreya; Cherian, Anoop

TR2023-154 December 27, 2023

Abstract

Audio-visual navigation of an agent towards locating an audio goal is a challenging task especially when the audio is sporadic or the environment noisy. In this paper, we present CAVEN, a Conversation-based Audio-Visual Embodied Navigation framework in which the agent may interact with a human/oracle for solving the task of navigating to an audio goal. Specifically, CAVEN is modeled as a budget-aware partially observable semi-Markov decision process that implicitly learns the uncertainty in the audio-based navigation pol- icy to decide when and how the agent may interact with the oracle. Our CAVEN agent can engage in fully-bidirectional natural language conversations by producing relevant questions and interpret free-form, potentially noisy responses from the oracle based on the audio-visual context. To enable such a capability, CAVEN is equipped with: i) a trajectory forecasting network that is grounded in audio-visual cues to produce a potential trajectory to the estimated goal, and (ii) a natural language based question generation and reasoning network to pose an interactive question to the oracle or interpret the oracle's response to produce navigation instructions. To train the interactive modules, we present a large scale dataset: AVN-Instruct, based on the Landmark-RxR dataset. To substantiate the usefulness of conversations, we present experiments on the benchmark audio-goal task using the SoundSpaces simulator under various noisy settings. Our results reveal that our fully-conversational approach leads to nearly an order-of-magnitude improvement in success rate, especially in localizing new sound sources and against methods that use only uni-directional interaction.

AAAI Conference on Artificial Intelligence 2023

CAVEN: An Embodied Conversational Agent for Efficient Audio-Visual Navigation in Noisy Environments

Xiulong Liu[1*], **Sudipta Paul**[2†], **Moitreya Chatterjee**[3], **Anoop Cherian**[3]

[1]University of Washington, Seattle, WA
[2]Samsung Research America, Mountain View, CA
[3]Mitsubishi Electric Research Labs, Cambridge, MA
xl1995@uw.edu, spaul007@ucr.edu, chatterjee@merl.com, cherian@merl.com

Abstract

Audio-visual navigation of an agent towards locating an audio goal is a challenging task especially when the audio is sporadic or the environment noisy. In this paper, we present CAVEN, a Conversation-based Audio-Visual Embodied Navigation framework in which the agent may interact with a human/oracle for solving the task of navigating to an audio goal. Specifically, CAVEN is modeled as a budget-aware partially observable semi-Markov decision process that implicitly learns the uncertainty in the audio-based navigation policy to decide when and how the agent may interact with the oracle. Our CAVEN agent can engage in fully-bidirectional natural language conversations by producing relevant questions and interpret free-form, potentially noisy responses from the oracle based on the audio-visual context. To enable such a capability, CAVEN is equipped with: i) a trajectory forecasting network that is grounded in audio-visual cues to produce a potential trajectory to the estimated goal, and (ii) a natural language based question generation and reasoning network to pose an interactive question to the oracle or interpret the oracle's response to produce navigation instructions. To train the interactive modules, we present a large scale dataset: AVN-Instruct, based on the Landmark-RxR dataset. To substantiate the usefulness of conversations, we present experiments on the benchmark audio-goal task using the SoundSpaces simulator under various noisy settings. Our results reveal that our fully-conversational approach leads to nearly an order-of-magnitude improvement in success rate, especially in localizing new sound sources and against methods that use only uni-directional interaction.

Introduction

The advent of powerful deep neural networks and sophisticated language models have led to significant advancements in building conversational agents that can collaborate with humans in solving challenging reasoning tasks (Peng et al. 2023; Ram et al. 2018; Chowdhery et al. 2022; Gupta and Kembhavi 2023; You et al. 2022). While, much effort has been expended on tasks that are predominantly in the language domain, such progress is yet to percolate into real world problems that need complex reasoning over multiple modalities of perception (Li et al. 2022; Liu et al. 2023). One such task that we exclusively explore in this paper is that of audio-visual navigation of an embodied robotic agent where the goal is to localize a sound producing source in a realistic, complex, and never-seen before environment when the sound is noisy, intermittent, sporadic, and mixed with other sounds — a situation even humans may find hard to tackle. As can be easily imagined, the applications needing such an *audio goal* capability are enormous; for example, at one end, we may think of a robotic disaster and emergency response agent that may need to move through huge rubble to localize victims who may cry for help and on the other, one may consider a home robotic vacuum repurposed to be vigilant to strange sounds.

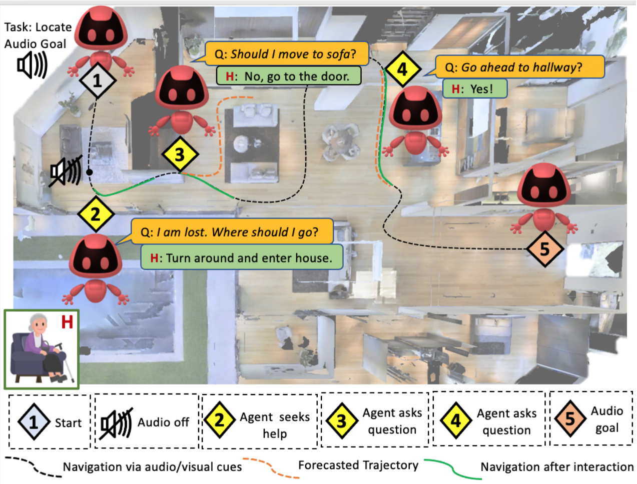

Figure 1: An illustrative CAVEN interaction: The agent starts at ◇1 guided by the audio event at ◇5. At ◇2, the agent decides to seek help from the human/oracle H (e.g., because the audio stopped). The oracle then provides a short natural language instruction for the agent to follow. At locations ◇3 and ◇4, the agent decides to ask questions to the oracle using the forecasted trajectories (orange) and gets feedback, finally reaching the audio goal at ◇5.

While the task of navigating to the audio goal, has witnessed some attention in the recent years (Chen et al., 2021), we consider a variant of this task, dubbed audio-visual-language embodied navigation (AVLEN) (Paul et al., 2022), where the agent has the ability to interact with a human/oracle when it is unable to solve the task by itself and potentially

[*] Work done while interning at MERL.
[†] Work done while the author was at UC Riverside.

query an oracle for task-specific guidance. However, the interactive abilities of the agent in AVLEN is limited in several aspects. In particular, the AVLEN agent could ask only a fixed question (e.g., “Help me!”), while the (human) oracle could provide a natural language response for guiding the agent to the goal. This technique of querying, while useful to some extent, does not cover the full scope of bi-directional interactions. As we know, back and forth interaction in natural language simulates a human-like conversation, allowing for better expressivity towards sharing ideas effectively. For instance, let’s assume for a moment that the agent is a 5 year old child who needs help in finding a sounding toy at a secret location. While the parents (oracle) could suggest: “look inside the wooden trunk” (as in Paul et al., 2022), the child might not know what a ‘trunk’ is. Instead, isn’t it better if the child had asked: ”Should I look next to the large brown box?” and the parents say: ”yes”? or suggest ”No, look inside it”? It is not only easy to respond with a ‘yes’/‘no’ answer (if possible), but this also avoids the need to know what a ‘trunk’ is (and ask more questions or make wrong inferences). Engaging in conversations to resolve such ambiguities is of importance in several time-critical real-world circumstances, e.g., the sound of wheezing in an elderly care or a thud in a medical facility.

Our goal in this paper is to build a fully-conversational robotic agent, which we call CAVEN – Conversational Audio-Visual Embodied Navigation, with the capabilities as described above, that can engage in bidirectional interactions with an oracle in natural language towards solving the audio goal task in a complex realistic visual environment. Specifically, CAVEN can either use the audio-visual cues for its navigation (as in prior works (Chen et al. 2020, 2021a; Gan et al. 2020)) or in case the agent is uncertain of which navigation step to take, it can interact with the oracle in two distinct modes: (i) a *question mode*, in which the agent forecasts a plausible trajectory based on audio-goal belief, using which it frames a natural language question to be posed to the oracle, and subsequently interpreting the oracle’s response to the question, and (ii) a *query mode*, where the agent is unsure of what question to even phrase (e.g., when there are no useful cues in the scene) or completely uncertain about its current situation, and therefore directly seeks the oracle’s guidance. Figure 1 illustrates a typical conversation between a human and our agent.

There are several challenges to tackle when designing the learning and inference model for CAVEN. Specifically, (i) when should the agent use language? (ii) what type of language interaction should the agent use (question or query)? (iii) how should the agent phrase the question? (iv) how to make the oracle understand the agent’s question?, (v) how should the oracle respond to the agent’s question? and (vi) how frequently should the agent be allowed to ask questions (budget)? Note that, some of these challenges are partially addressed in prior works (Kesiraju et al. 2020; Siddhant and Lipton 2018; Xiao and Wang 2019) such as (v) and (vi). However in CAVEN, we tackle all these challenges within a single framework, by proposing a novel budget-aware partially observable semi-Markov decision process (POSMDP), using a reinforcement learning framework by introducing novel learning rewards.

To empirically assess the performance of CAVEN, we conduct extensive experiments on the semantic audio-goal navigation task (Chen et al. 2021b) in the SoundSpaces simulator, under various challenging scenarios, each having intermittent sounds emanating from a source. One key difficulty to train the CAVEN model is the absence of any large scale dataset that includes language instructions in an audio-visual navigation setting. To this end, we introduce *AVN-Instruct* – a novel audio-visual-language navigation sub-instruction dataset with 41.5k pairs of audio-goal, trajectory, and language instructions. Our experimental results using the above setup clearly bring out the benefits of enabling the agent to converse with the oracle, demonstrating a solid gain of nearly 12% over competing approaches on the success rate.

We summarize below the core contributions of our work:

- We present CAVEN, a multimodal navigation agent that is, for the first time, capable of fully-bidirectional interaction with an oracle in free-form natural language, thereby facilitating easy communication.
- We introduce a novel *question module* for bi-directional interaction with the oracle consisting of: (i) a trajectory forecasting module grounded on both visual scenes and audio cues, (ii) a question generation module, and (iii) a question decoder (FollowerNet, on the oracle).
- We design a novel budget-aware and uncertainty-splitting reinforcement learning policy, which integrates the question module as additional policy (using suitable reward design inspired by differential RL) in addition to audio-visual navigation and language-based policy.
- We propose a novel audio-visual-language navigation sub-instruction dataset, AVN-Instruct to pre-train embodied navigation models. We also propose two new metrics to evaluate language-guided navigation tasks, dubbed SNO and SNI.
- Our experiments demonstrate state-of-the-art performances against related prior approaches by an order-of-magnitude increase in success rate.

Related Works

Audio-Visual Embodied Navigation Tasks: Recent years have seen several works in Embodied AI that consider the audio-goal navigation task (Chen et al. 2020, 2021a; Gan et al. 2020; Yu et al. 2022). Generally, this task assumes a continuous sound. However, there are derivatives that look at situations when the audio is sporadic and depends on the category of the sounding object, dubbed semantic audio-goal navigation (Chen et al. 2021b). Both of these tasks are facilitated by the SoundSpaces simulator (Chen et al. 2020) that can render realistic audio in 3D visual environments. While the aforementioned methods only consider audio and visual modalities, (Paul et al., 2022) proposes AVLEN that utilizes language feedback from the oracle. However, there is no provision of posing questions, which burdens the oracle with the task of chalking out a path to the goal whenever help is sought. Contrary to these approaches, our proposed

CAVEN utilizes bi-directional interaction with the oracle besides audio-visual cues, a setting that is more practical.

Vision-and-Language Navigation (VLN): The task in VLN is to use (or execute) natural language instructions to reach a target location. Akin to (Gu et al. 2022), we group VLN approaches in three categories: (i) instruction-at-start, (ii) oracle guidance, and (iii) bi-directional interaction. *Instruction-at-start* is a well-explored research area (Anderson et al. 2018; Hong et al. 2021; Ke et al. 2019; Liu et al. 2021; Majumdar et al. 2020; Ma et al. 2019a,b; Zhu et al. 2020; Chen et al. 2021c; Pashevich et al. 2021; Guhur et al. 2021) in which the agent is given a language instruction at its start describing the intended path. To tackle the task, Wang et al. (Wang et al. 2019) uses cross-modal attention to focus on the relevant parts of both vision and language modalities, while others (Fried et al. 2018; Tan et al. 2019), used augmented instruction-trajectory pairs to improve the VLN performance. Recent approaches have begun using transformer-based architectures, such as BERT (Devlin et al. 2018) for VLN (Hong et al. 2021; Majumdar et al. 2020). In the *oracle guidance* setting, an agent may receive feedback (ground truth actions (Chi et al. 2020), encoded ground truth action (Nguyen et al. 2019a), or a fixed set of natural language instructions (Nguyen et al. 2019b)) from an oracle during navigation. A major challenge in these works, however, is to identify when to query an oracle for feedback. In the *bi-directional interaction* setting, an agent can use natural language to seek navigation help (Banerjee et al. 2021; Thomason et al. 2020; Cao et al. 2022; Lin et al. 2022). Thomason et al. (Thomason et al. 2020) introduced the CVDN dataset with human-human dialogue for navigation. However, these works allow the agent and oracle to communicate only at certain locations of the environment, making it less practical to real world scenarios. Self-Motivated Communication Agent (SCoA) (Zhu et al. 2021) permits the agent to only ask templated questions filled in with labels of detected scene objects, grossly limiting the nature of interaction between the agent and the oracle. Contrary to these methods, we empower our CAVEN agent with: (i) the ability to seek occasional human/oracle help at any location and (ii) competence for natural language-based scene grounded conversations with an oracle for effective navigation.. Further, our agent is also robust to noisy feedback from the oracle.

LLM-based Embodied Navigation. The spark of recent advancements in large language models (LLMs) (Bubeck et al. 2023; Touvron et al. 2023; OpenAI 2023) has brought along new opportunities in improving multi-modal robot navigation. In the context of Vision and Language Navigation, early works like LM-Nav (Shah et al. 2022) analyzed landmarks in the instruction to be used for visual navigation. In NavGPT, (Zhou et al. 2023) explored the possibility of integrating ChatGPT (Ouyang et al. 2022) with a vision foundation model: BLIP-2 (Li et al. 2023) into its prompting setup to perform multi-modal reasoning to navigate in a zero-shot manner. While, these works achieve decent performances on vision-language navigation task, they do not incorporate audio as part of the inputs and are thus complementary to our efforts.

Proposed Method

Task Setup: We assume the standard embodied audio goal problem setup (Chen et al. 2020), where the agent is equipped with an RGBD camera and a binaural microphone and at any time step can take one of four navigation actions: $\{\mathrm{stop}, \mathrm{move\_forward}, \mathrm{turn\_right}, \mathrm{turn\_left}\}$ in a densely-sampled 3D grid with the goal of locating the audio source. As in (Chen et al. 2020), we assume the sound is semantically unique and is produced by a static object, however the audio could be noisy, sporadic, or mixed up with other environmental sounds. An audio goal navigation episode is deemed successful if the agent calls the stop action within a given proximity to the goal.

Beyond the standard problem setup above, our CAVEN agent can also seek language-based guidance from an oracle. Practically, the oracle could be a human who has higher level information about the scene, e.g., a remote operator controlling several such agents and intervening whenever needed, or a home owner who is notified about the situation and is sought to provide guidance. To incorporate the language modality into the audio goal setup, we follow AVLEN (Paul, Roy-Chowdhury, and Cherian 2022) in which the agent can query the oracle for help and the oracle responds via a short message describing a pathlet towards the audio goal. However as is clear, the interaction in AVLEN is only uni-directional and the agent cannot ask questions. Our CAVEN agent goes beyond this shortcoming and can phrase a question in free-form natural language using cues from the audio-visual context. Further, we assume the oracle after receiving this question, will either give a *"yes"* response if the oracle's interpretation of the question in its own state space results in actions that match its estimate of the actions along the ground truth geodesic to the goal. Otherwise, the oracle responds with a *"no"* followed by a short sentence guiding the agent to the goal. Note that the oracle in AVLEN has access to the 3D space of the full environment and thus can provide plausible instructions for navigation, however the CAVEN agent has only a *very restricted view* of the scene in its vicinity, thus making this task of creating a question at the agent's side entirely different from that of the oracle's. In our new problem setup, we also assume that the number of times an agent can receive direct navigation instructions from the oracle (as a result of a wrong question or when it directly queries) is limited by a budget so that the agent only seeks help when necessary.

CAVEN Learning and Inference Framework

As we envisage CAVEN to incorporate various modules with diverse temporal spans, it is natural to consider a partially observable semi-Markov decision process (POSMDP) as our control module (Le et al. 2018). A POSMDP is essentially a partially observable Markov decision process (POMDP) with macro actions and is characterized by the tuple $(\mathcal{S}, \mathcal{A}, T, R, \Omega, \mathcal{Z}, \gamma)$ where $\mathcal{S}$, $\mathcal{A}$, T, R, and γ are the state space, action space, transition function, reward function, and discount factor, respectively, while Ω and $\mathcal{Z}$ are the observation space and observation model. In a partially observable setup, the agent maintains a belief distribution

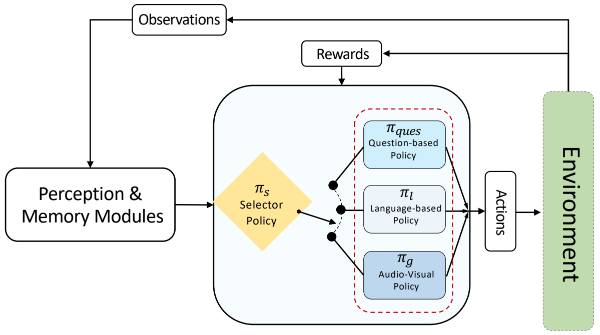
Observations
Rewards
Perception & Memory Modules
π_s Selector Policy
π_{ques} Question-based Policy
π_l Language-based Policy
π_g Audio-Visual Policy
Actions
Environment

Figure 2: Architecture of our CAVEN model. We show the reinforcement learning policies, namely a selector policy π_s and three option policies π_g, π_l, and π_{ques}.

b over $\mathcal{S}$, which is used to compute the expected reward. While in a POMDP setup, the agent maintains a policy $\pi : \mathbb{R}^{|\mathcal{S}|} \times \mathcal{A} \rightarrow [0, 1]$ that maximizes the expected reward, in POSMDP the agent maintains multiple low level 'options' as temporal abstractions, denoted Ξ, and a high level selector policy π_s to select the options from Ξ. An option $\xi \in \Xi$ is defined by the triplet $(\mathcal{S}^{\xi}, \pi^{\xi}, \beta^{\xi})$, where $\mathcal{S}^{\xi}$ is the set of valid states, π^{ξ} is the policy, β^{ξ} is the termination condition. In our setup, we disentangle agents' interactive audio goal navigation process into three low level temporal abstractions (i.e., options): i) audio-visual navigation ξ_g, ii) instruction-guided navigation ξ_l, iii) bi-directional question-answer navigation ξ_{ques}. We use π_g, π_ℓ, and π_{ques} to denote the respective policies of ξ_g, ξ_l, and ξ_{ques} and R'_g, R'_l, and R'_{ques} as corresponding immediate rewards. In our case, instead of using the termination condition for each option, we allow the audio-visual navigation option ξ_g to take a single step, while the interaction-based options (ξ_l and ξ_{ques}) are allowed a fixed span of ν steps (unless stop action is executed by these options). These options are assumed valid in any state of the environment, i.e., $\mathcal{S}_{\xi_g}, \mathcal{S}_{\xi_l}, \mathcal{S}_{\xi_{ques}} \in \mathcal{S}$.

Although the agent always has access to three option policies, it should maintain its autonomy and should only engage in a limited number of language interactions to mitigate its uncertainty. Further, in our setup, we have different levels of engagement of the oracle with the agent for varied language interactions (e.g., bi-directional conversations with question and answer, querying for language instructions) and a system should favor asking correct questions based on its audio-visual cues over relying on oracle instructions to reduce the oracle's effort. To consider all of these scenarios, we formulate option policies with dynamically adjusted constraints. These constraints are realized by penalties associated with the reward functions of each option policy. The audio-visual navigation policy $\pi_g : \mathbb{R}^{|\mathcal{S}| \times |M|} \times G \times |\mathcal{A}| \rightarrow [0, 1]$ chooses the navigation actions $a \in \mathcal{A}$ based on the audio-visual features. Here, M is a memory module storing a fixed number of past observations, and G is a set of audio goal estimates. Since, π_g is fully autonomous and does not require oracle interaction, we encourage selecting this option by defining an unconstrained reward, $R'_g(b_t, a_t) = \mathbb{E}\left[\sum_{i=t}^{\infty} \gamma^{i-t} R'_g(b_i, a_i)\right]$. The instruction guided navigation policy $\pi_\ell : \mathbb{R}^{|\mathcal{S}| \times \nu} \times \mathcal{I} \times G \times |\mathcal{A}| \rightarrow [0, 1]$ navigates based on the received natural language instruction. Here, $\mathcal{I}$ is the set of all natural language instructions. Since, π_ℓ is entirely dependent on the oracle instruction, we penalize such interactions using ζ_ℓ, i.e., $R'_\ell(b_t, a_t) = \mathbb{E}\left[\sum_{i=t}^{t+\nu-1} \gamma^{i-t} R'_\ell(b_i, a_i)\right] - \zeta_\ell(t)$. The bi-directional conversational navigation policy $\pi_{ques} : \mathbb{R}^{|\mathcal{S}| \times \nu} \times \mathcal{Q} \times \mathcal{I} \times G \times |\mathcal{A}| \rightarrow [0, 1]$ navigates based on asking a question and receiving an answer. Here, $\mathcal{Q}$ is the set of all natural language questions. Specifically, π_{ques} consists of multiple novel components and the policy module can be divided in three submodules based on the functionality: i) question generator $\mathcal{G}^q$, ii) question evaluator $\mathcal{E}$, and iii) instruction generator $\mathcal{G}^i$. The output of π_{ques} depends on the interplay between these submodules. Question generator $\mathcal{G}^q$ is used to generate questions. Then, the question evaluator $\mathcal{E}$ evaluates on the oracle side if the question is correct. If the question is incorrect then the instruction generator $\mathcal{G}^i$ (which mimics the oracle) generates instructions for navigation. Since, asking correct question results in minimal oracle effort in producing a response, we define a dynamic penalty based on the question by, $R'_{ques}(b_t, a_t) = \mathbb{E}\left[\sum_{i=t}^{t+\nu-1} \gamma^{i-t} R'_{ques}(b_i, a_i)\right] - \zeta_{ques}(t, \mathcal{E}(q))$, where $q \in \mathcal{Q}$ and $\mathcal{E}(q)$ is an indicator function that checks whether the question q asked by the agent falls within the range of the estimated navigation direction by the oracle, and no penalty will incur when $\mathcal{E}(q) = 1$.

Putting it all together, our objective to learn these policies $\pi = \{\pi_s, \pi_g, \pi_\ell, \pi_{ques}\}$ is via maximizing the value function $V^{\pi}(b_0)$, i.e.,

$$\arg\max_{\pi} V^{\pi}(b_0), \text{ where}$$

$$\begin{aligned} V^{\pi}(b) = \pi_s(\xi_g|b) \left[R'_g + \sum_{o' \in \Omega} \mathcal{Z}'(o'|b, \xi_g) V^{\pi}(b') \right] \\ + \pi_s(\xi_\ell|b) \left[R'_\ell + \sum_{o' \in \Omega} \mathcal{Z}(o'|b, \xi_\ell) V^{\pi}(b') \right] \\ + \pi_s(\xi_{ques}|b) \left[R'_{ques} + \sum_{o' \in \Omega} \mathcal{Z}'(o'|b, \xi_{ques}) V^{\pi}(b') \right]. \end{aligned} \quad (1)$$

Here, b' is the updated belief and $\mathcal{Z}'$ is the multi-time transition function (Sutton, Precup, and Singh 1999) given by: $\mathcal{Z}'(o'|b, \xi) = \sum_{j=1}^{\infty} \sum_{s'} \sum_{s} \gamma^j \mathcal{Z}(s', o', j|s, \xi) b(s)$. Below, we detail the architecture of each of these policies.

Bi-directional Question-Answer Policy Module: Bi-directional question-answer policy consists of three components: (i) *TrajectoryNet* (forecasting short navigation steps), (ii) *QuestionNet* (generates natural language questions using trajectories), and (iii) *FollowerNet* (interprets the question on oracle side). These components detailed below are illustrated in Figure 3. They are used to enable the functionalities within the π_{ques} policy as: i) question generator (*TrajectoryNet* + *QuestionNet*), ii) question evaluator (*FollowerNet*), and iii) instruction generator (*QuestionNet*).

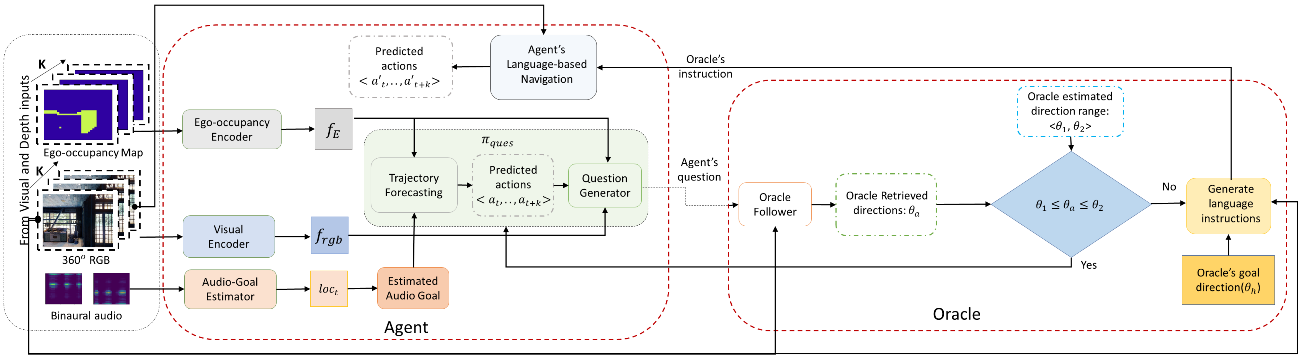
From Visual and Depth inputs
K
Ego-occupancy Map
K
360^o RGB
Binaural audio
Predicted actions $< a'_t, .., a'_{t+k} >$
Agent's Language-based Navigation
Ego-occupancy Encoder
f_E
π_{ques}
Trajectory Forecasting
Predicted actions $< a_t, .., a_{t+k} >$
Question Generator
Visual Encoder
f_{rgb}
Audio-Goal Estimator
loc_t
Estimated Audio Goal
Agent
Oracle's instruction
Agent's question
Oracle estimated direction range: $<\theta_1, \theta_2>$
Oracle Follower
Oracle Retrieved directions: θ_a
$\theta_1 \le \theta_a \le \theta_2$
No
Yes
Generate language instructions
Oracle's goal direction(θ_h)
Oracle

Figure 3: Architecture of our question policy module and the control flow within it. Here, θ_a is oracle-interpreted agent's direction to take, while θ_1 and θ_2 represent the lower and upper bounds of oracle's estimated direction range to the goal.

(i) TrajectoryNet: In order to forecast the steps of a trajectory, the agent needs to have a clear observation of its surroundings. Towards this end, we allow the agent to have a panoramic view at its current location. With the full view of its surroundings and an estimate of the audio-goal, the agent *forecasts* a sequence of l-step actions, denoted by $\mathcal{F}_a$. This is achieved by *TrajectoryNet* – a transformer encoder-decoder network which takes as input a sequence of ego occupancy maps E_t of four disjoint views (separated by 90-degrees) and the goal vector g_t predicted by a binaural audio encoder, to predict a sequence of actions $\mathcal{F}_a = \langle f_{a_1}, f_{a_2} \dots, f_{a_l} \rangle$ (auto-regressively). The ego occupancy map is calculated by transforming depth images into point clouds and projecting them onto the ground plane.

(ii) QuestionNet: The action sequences defined in SoundSpaces (Chen et al. 2020) are discrete, e.g, $\mathrm{move\_forward}$ implies moving forward by $1m$. However, the language produced from these actions by itself may be ambiguous (since it is a higher level construct) and thus does not explicitly reflect the granularity of these discrete actions. Further, as will be explained in the Experiments section, while the trajectories are forecasted using the SoundSpaces grid (which uses 90 degree angles for turning), the language instructions are produced using a model trained on the LandmarkRxR dataset (He et al. 2021), that uses panoramic images as input. To compensate for these mismatches, we propose to first gather the view of the agent at the end of the forecasted trajectory, which we call g_{view}, and the corresponding displacement vector $g_{sub} := [d_f, cos(\theta_f), sin(\theta_f)]$, where d_f is the distance between the agent's location and the trajectory end point and θ_f is the angle difference between the direction of d_f and the agent's facing direction.

Next, we capture the panorama around the agent using 12 equiangular views, as RGB images as well as the corresponding occupancy maps to abstract the 3D scene geometry. ResNet-152 features are then extracted from these RGB images using an ImageNet pre-trained model, while the ego-occupancy maps are encoded using a 2D-CNN; both the features are fused with position embeddings and passed through a transformer encoder. In order to fuse these panoramic views with the forecasted agent views (in the SoundSpaces grid), we propose to use a transformer decoder, which takes the output of the encoder and a fusion of ResNet-152 features from g_{view}, coupled with the position encoding of g_{sub}, and the embeddings of hitherto produced words in the question (e.g., GloVe (Pennington, Socher, and Manning 2014) or CLIP (Radford et al. 2021)), and proceeds to generate the next word in the question autoregressively.

(iii) FollowerNet: After the question is asked, the oracle needs to verify if it can be correctly translated into a direction that falls within the oracle's own estimation of the direction range to the goal. To this end, we incorporate *FollowerNet* at the oracle, which is assumed to have knowledge of the agent's location and its audio-visual context, and can convert the question back to the oracle's space of the view angles. See Appendix for details on its training.

Language-based Policy Module: There can be situations when an agent cannot produce a question to ask the oracle; e.g., when there are no useful landmarks to base the question. To cater to such cases, we equip the agent to directly query the oracle for language-based instructions. When invoked, the agent receives instructions, similar to when a wrong question is posed to the oracle.

Audio-Visual Navigation Policy Module: This policy is modeled as a transformer (Vaswani et al. 2017) based encoder-decoder as in (Chen et al. 2021b). The encoder takes as input the current and previous observations in the memory M, the output of which is combined with the goal descriptor g and decoded by the decoder to produce a feature vector defining the belief state of the agent b. Next, a single-layer actor-critic neural network learns a policy, π_g, that transforms this belief b to predict the distribution on the navigation actions, which the agent samples to take a step in the environment.

Selector Policy: This module, denoted π_s, decides when to navigate using audio-visual cues (i.e., use π_g), when to query the oracle for instructions directly (i.e., use π_ℓ); or when to pose a question to the oracle, (i.e., use π_{ques}). Instead of directly using model uncertainty (as is common in prior works (Chi et al. 2020)), we use our proposed RL framework to train this policy in an end-to-end manner, guided by the reward design ζ described below.

Table 1: Comparison of CAVEN performances against the state of the art under heard and unheard sound settings.

		Heard Sound							Unheard Sound						
	Feedback	**Success ↑**	**SPL ↑**	**SNA ↑**	**DTG ↓**	**SWS ↑**	**SNI ↑**	**SNO ↑**	**Success ↑**	**SPL ↑**	**SNA ↑**	**DTG ↓**	**SWS ↑**	**SNI ↑**	**SNO ↑**
Random Nav.	✗	1.4	3.5	1.2	17.0	1.4	-	-	1.4	3.5	1.2	17.0	1.4	-	-
ObjectGoal RL	✗	1.5	0.8	0.6	16.7	1.1	-	-	1.5	0.8	0.6	16.7	1.1	-	-
Gan et al. (Gan et al. 2020)	✗	29.3	23.7	23.0	11.3	14.4	-	-	15.9	12.3	11.6	12.7	8.0	-	-
Chen et al. (Chen et al. 2020)	✗	21.6	15.1	12.1	11.2	10.7	-	-	18.0	13.4	12.9	12.9	6.9	-	-
AV-WaN (Chen et al. 2021a)	✗	20.9	16.8	16.2	10.3	8.3	-	-	17.2	13.2	12.7	11.0	6.9	-	-
SMT(Fang et al. 2019)+Audio	✗	22.0	16.8	16.0	12.4	8.7	-	-	16.7	11.9	10.0	12.1	8.5	-	-
SAVi (Chen et al., 2021)	✗	33.9	24.0	18.3	8.8	21.5	-	-	24.8	17.2	13.2	9.9	14.7	-	-
AVLEN (Paul et al., 2022)	Language	36.1	24.6	19.7	8.5	23.1	-	21.8	26.2	17.6	14.2	9.2	15.8	-	15.9
AVLEN (Paul et al., 2022)	GT Actions	48.2	34.3	26.7	7.5	36.0	-	29.1	36.7	24.1	18.7	8.3	26.6	-	22.3
CAVEN (Ours)	Noisy-Language	**45.2**	**32.9**	**28.8**	**7.5**	**32.3**	17.9	**31.4**	**38.2**	**27.6**	**24.1**	**8.2**	**25.9**	15.0	**23.1**
CAVEN (Ours)	Language	**48.4**	**35.8**	**31.0**	**6.9**	**34.2**	21.5	**33.4**	**42.0**	**30.0**	**26.5**	**7.6**	**30.9**	16.7	**27.9**
CAVEN (Ours)	GT Actions	**54.8**	**41.4**	**35.9**	**6.5**	**39.9**	24.3	**37.8**	**49.7**	**37.3**	**32.7**	**6.7**	**37.2**	19.8	**33.0**

Reward Design

In this section, we detail the rewards structure to train the various policy modules in an end-to-end manner. For the π_g policy, we use the reward scheme in (Chen et al. 2020), i.e., the agent gets $+1$ for moving towards the goal and receives $+10$ if it calls the stop near the goal. Further, to make the navigation efficient, we penalize by -0.01 for every step taken. The penalty structure for the language-based policies is designed so as to discourage the agent to seek help from the oracle, while also limiting the number of instructions $K \geq 0$ received. To this end, we propose a dynamic penalty that increases in magnitude as more instructions are sought from the oracle. Specifically, if $\zeta_l(k, K)$ denotes the penalty received by the agent for the k-th query, then

$$\zeta_l(k, K) = \begin{cases} \frac{k \times (r_{neg} + \exp(-\nu))}{\nu} & k < K \\ r_{neg} + \exp(-k) & k \geq K, \end{cases} \quad (2)$$

where ν characterizes the number of steps agent takes based on the language instruction received, which is fixed in our case, and $r_{neg} = -0.6$ is a constant. Until $k < K$, the penalty is linear, however for $k \geq K$, the penalty approaches r_{neg} exponentially thereby discouraging the agent to seek language guidance directly. Further to this penalty, we also include an additional cost for seeking oracle guidance frequently. Specifically, we include a linear penalty ζ_f if the agent queries the oracle within τ steps, where $\zeta_f(j) = \frac{r_f}{j}$ for the j-th step, if $j \in [0, \tau]$ and zero otherwise (with $r_f = -0.5$). Thus, the total penalty for the agent querying the oracle is given by $\zeta_l + \zeta_f$.

As the question policy π_{ques} blends between π_g and π_ℓ, we propose a penalty structure that integrates both these policies. Specifically, if $\zeta_{ques}(m)$ is the penalty incurred by the agent for asking the m-th question, then

$$\zeta_{ques}(m) = \zeta_l(m, K')\, \delta_{ques}(m) + \zeta_{f_{ques}}(m), \quad (3)$$

where $\zeta_{f_{ques}}$ is the penalty for asking questions too many times (similar to $\zeta_f(k)$), K' is the budget on the number of wrong questions, and $\delta_{ques}(m) = 1$ if the response to the m-th question by the oracle is 'no', else $\delta_{ques}(m) \in [0, 1)$ is a constant. In our experiments, we find that not penalizing the agent for correct questions leads to better results, i.e., $\delta_{ques}(m) = 0$. Such a differential reward implicitly reinforces the agent to learn correct trajectories to the audio goal, improving performance. We also couple π_{ques} with π_ℓ via enforcing $K + K' = \eta$ for an $\eta = 3$. Using this reward setup, the policies are trained with the DD-PPO algorithm (Wijmans et al. 2019).

Experiments

Datasets: The CAVEN agent is trained and evaluated on the SoundSpaces platform (Chen et al. 2020). It uses MatterPort3D environment scans (Chang et al. 2017). We use the the semantic audio-visual navigation dataset from (Chen et al. 2020) to benchmark our experiments. The details of the dataset are provided in the Appendix.

AVN-Instruct Dataset: For pre-training and evaluation of the language interaction modules (i.e., *QuestionNet*, *FollowerNet*), we use the Landmark RxR dataset (He et al. 2021), which contains 150k well-annotated sub-trajectories and their corresponding language sub-instructions grounded on scenes captured using the MatterPort3D simulator. Then we adopt the pre-trained QuestionNet to synthesize a dataset called AVN-Instruct, which contains a total of 41.5k dense pairs of sub-instructions, audio-goal, and visual scene under the state space of Soundspace Habitat simulator, by sampling the trajectories and transporting the grid from Matterport3D to Soundspace and obtaining the action sequence which closely approximates this trajectory. Before integrating the modules into the RL framework, we fine-tune the whole question module end-to-end on AVN-Instruct with a set of 500 and 1000 samples for validation and testing.

Evaluation Metrics: We follow the standard metrics defined in SAVi (Chen et al. 2021b) to evaluate the navigation performance, namely: (i) success rate (SR) for navigation success, (ii) success rate weighted by inverse path length (SPL), (iii) success rate weighted by inverse number of actions (SNA), (iv) average distance to goal (DTG), and (v) success rate when silent (SWS). In addition, we introduce two new metrics for assessing navigation performance that also considers the number of language-based oracle interactions, namely: (a) success rate weighted by the inverse number of language interactions (SNI) – which is the ratio of the success rate to the average total number of times either direct instructions are sought from the oracle or a question is posed to it (averaged by the number of episodes), and (b) success rate weighted by inverse number of oracle instructions (SNO) – which is the ratio of the success rate to the average total number of times either direct instructions are sought from the oracle or a *wrong* question is posed to it. These additional metrics help explain the performance gain under conversational settings.

Experimental Results and Analysis: Here, we compare our proposed formulation against state-of-the-art semantic audio-visual navigation approaches, namely (Gan et al. 2020), (Chen et al. 2020), AV-WaN (Chen et al., 2021),

Table 2: Comparison of CAVEN performances with different approaches in the *presence of distractor sound.*

	Feedback	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑	SNI ↑	SNO ↑
Chen et al	✗	4.0	2.4	2.0	14.7	2.3	-	-
AV-WaN (Chen et al. 2021a)	✗	3.0	2.0	1.8	14.0	1.6	-	-
SMT+Audio (Fang et al. 2019)	✗	4.2	2.9	2.1	14.9	2.8	-	-
SAVi (Chen et al., 2021)	✗	11.8	7.4	5.0	13.1	8.4	-	-
AVLEN (Paul et al., 2022)	Language	14.0	8.4	5.9	12.8	11.1	-	8.5
Random	Bi-interact	16.9	10.6	7.9	11.9	11.1	7.2	9.4
Uniform	Bi-interact	16.9	10.5	7.6	11.9	11.6	7.1	9.5
Model Uncertainty	Bi-interact	19.6	12.4	8.9	11.4	14.0	7.8	10.2
CAVEN	Bi-interact	**21.3**	**13.9**	**11.7**	**11.6**	**14.5**	8.4	**11.6**

Table 3: Ablation of the reward parameter δ_{ques} of CAVEN's question module under unheard sound settings.

Architecture	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑
CAVEN (δ_{ques}=1.0)	32.1	23.1	19.4	8.0	20.8
CAVEN (δ_{ques}=0.5)	36.5	26.9	24.6	8.2	21.1
CAVEN (δ_{ques}=0.0) (ours)	**42.0**	**30.0**	**26.5**	**7.6**	**30.9**

SMT (Fang et al. 2019) + Audio, SAVi (Chen et al., 2021) and AVLEN (Paul et al., 2022). Using the same protocol as in AVLEN, we evaluate our performances on two different settings: (i) heard and (ii) unheard sound, both in unseen environments with sporadic sources. To ensure the comparisons are fair, we control our CAVEN model to have a similar number of oracle feedbacks as in Paul et al.. Table 1 provides the results of our experiments using heard and unheard sounds. The table shows that our full model –CAVEN (language), is capable of achieving significant improvements across all metrics. CAVEN exhibits a **12% gain** on the newly introduced **SNO** metric over Paul et al., our closest competitor, in both heard and unheard cases. This clearly shows that the agent benefits much more from both our novel language components. Given the budget on directly receiving instructions from the oracle, we find that CAVEN poses a correct question about 40% of the time, thereby incurring less penalty. Even with a noisy oracle, i.e., *Noisy-Language* in Table 1, we achieve better performances compared to Paul et al., showing the robustness of our framework. To induce noise, we either ground the generated oracle's instructions on random trajectories or switch 'yes'/'no' responses, both with a chance of 25%.

Navigation Under Distractor Sounds: We also evaluate the performance of CAVEN in the presence of distractor sounds, in the unheard setting. Since this environment presents a mixture of sounds, therefore to disambiguate, a one hot encoding of the target sounding object is also provided as an input to the agent (as is the standard evaluation protocol (Chen et al. 2021b)). The presence of distractor sounds adversely affects the estimation of the audio-goal, which results in more uncertainty in the agent's decision-making. Under this setting, the conversations between the agent and oracle becomes even more critical. Even under such challenging circumstances, as shown in Table 2, we notice a 5.5% and a 3.1% **gain** on SPL and SNO, respectively against our closest competitor.

Ablation on Selector Policy: In Tables 1, 2, we compare various strategies instead of learning the selector policy, π_s. In *Random*, the agent randomly selects a navigation policy, while in *Uniform*, the agent chooses a policy every 3 steps, alternating between the three policies. In Model Uncertainty, the audio-goal uncertainty estimated by the selector policy is used to decide which policy to invoke; i.e., if the audio-goal uncertainty is above 66.7%, the language-based policy is invoked; if the uncertainty is between 33.3% and 66.7%, question policy is invoked; otherwise, the audio-goal policy is invoked. Our results show learning of π_s is better.

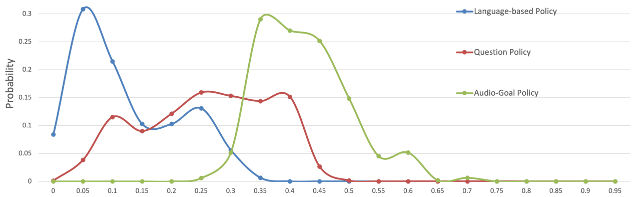

Figure 4: Distribution of estimated audio goal confidence when each policy is invoked.

Analysis of Policy Dynamics: To study the situations when the agent invokes the various navigation policies, we record the confidence of audio-goal estimated by selector policy π_s, when each of the option policies is invoked and compute its distribution using all test set episodes. As shown in Figure 4, the audio-goal is invoked when the agent is highly confident and the language-based policy is invoked when agent's confidence is low. It is note-worthy that the question policy is invoked more often when the agent is moderately confident. Though it potentially risks being penalized by asking wrong questions, it benefits from seeking confirmation from the oracle using its own audio-visual cues to help alleviate navigation uncertainty, thus facilitating efficient navigation.

Insights into Differential Rewarding: In Table 3, we report the CAVEN performances on varying the penalty parameter δ_{ques}. Note that our differential rewarding scheme gives no penalty when the agent makes a correct question $\delta_{ques} = 0$, however penalizes heavily for mistakes. Thus, the *gap* between the two penalties act as an incentive for the agent to make more number of correct trajectory predictions than in a case where this penalty gap is lower (e.g., $\delta_{ques} = 0.5, 1.0$ in which case it is similar to the penalty it receives for the wrong question). The success rate is much higher suggesting that the incentive the agent receives in making a correct question influences the learning of the trajectory forecasting significantly more.

Conclusions

In this paper, we introduced CAVEN for embodied navigation in an audio-visual setting for the audio goal task, where the agent is also equipped to converse with an oracle in natural language, when uncertain. We introduced a novel budget-aware partially observable semi-Markov decision process to learn the various control policies for solving the task. Quantitative evaluations of CAVEN under various noisy problem settings, using established and novel metrics, demonstrate large improvements in performance over competing methods, substantiating the benefits of our proposed interaction policies and our architecture. However, the interactions with the oracle might result in the agent having to wait for feedback, which we intend to fix in future work.

References

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and Van Den Hengel, A. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3674–3683.

Banerjee, S.; et al. 2021. The RobotSlang benchmark: Dialog-guided robot localization and navigation. In *Conference on Robot Learning*, 1384–1393. PMLR.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.

Cao, Y.; Lu, K.; DeFazio, D.; and Zhang, S. 2022. Goal-oriented Vision-and-Dialog Navigation via Reinforcement Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4473–4482.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

Chen, C.; et al. 2020. Soundspaces: Audio-visual navigation in 3d environments. In *European Conference on Computer Vision*, 17–36. Springer.

Chen, C.; et al. 2021a. Learning to Set Waypoints for Audio-Visual Navigation. In *International Conference on Learning Representations*.

Chen, C.; et al. 2021b. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15516–15525.

Chen, S.; Guhur, P.-L.; Schmid, C.; and Laptev, I. 2021c. History aware multimodal transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34.

Chi, T.-C.; Shen, M.; Eric, M.; Kim, S.; and Hakkani-tur, D. 2020. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2459–2466.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fang, K.; Toshev, A.; Fei-Fei, L.; and Savarese, S. 2019. Scene memory transformer for embodied agents in long-horizon tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 538–547.

Fried, D.; Hu, R.; Cirik, V.; Rohrbach, A.; Andreas, J.; Morency, L.-P.; Berg-Kirkpatrick, T.; Saenko, K.; Klein, D.; and Darrell, T. 2018. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31.

Gan, C.; Zhang, Y.; Wu, J.; Gong, B.; and Tenenbaum, J. B. 2020. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 9701–9707. IEEE.

Gu, J.; Stefani, E.; Wu, Q.; Thomason, J.; and Wang, X. E. 2022. Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions. *arXiv preprint arXiv:2203.12667*.

Guhur, P.-L.; Tapaswi, M.; Chen, S.; Laptev, I.; and Schmid, C. 2021. Airbert: In-domain pretraining for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1634–1643.

Gupta, T.; and Kembhavi, A. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14953–14962.

He, K.; et al. 2021. Landmark-RxR: Solving Vision-and-Language Navigation with Fine-Grained Alignment Supervision. In Ranzato, M.; and etal, eds., *Advances in Neural Information Processing Systems*, volume 34, 652–663. Curran Associates, Inc.

Hong, Y.; Wu, Q.; Qi, Y.; Rodriguez-Opazo, C.; and Gould, S. 2021. VLN BERT: A Recurrent Vision-and-Language BERT for Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1643–1653.

Ke, L.; Li, X.; Bisk, Y.; Holtzman, A.; Gan, Z.; Liu, J.; Gao, J.; Choi, Y.; and Srinivasa, S. 2019. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6741–6749.

Kesiraju, S.; Plchot, O.; Burget, L.; and Gangashetty, S. V. 2020. Learning document embeddings along with their uncertainties. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2319–2332.

Le, T. P.; et al. 2018. A deep hierarchical reinforcement learning algorithm in partially observable Markov decision processes. *Ieee Access*, 6: 49089–49102.

Li, G.; Wei, Y.; Tian, Y.; Xu, C.; Wen, J.-R.; and Hu, D. 2022. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. arXiv:2203.14072.

Li, J.; et al. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Lin, C.; Jiang, Y.; Cai, J.; Qu, L.; Haffari, G.; and Yuan, Z. 2022. Multimodal transformer with variable-length memory for vision-and-language navigation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, 380–397. Springer.

Liu, C.; Zhu, F.; Chang, X.; Liang, X.; Ge, Z.; and Shen, Y.-D. 2021. Vision-language navigation with random environmental mixup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1644–1654.

Liu, X.; et al. 2023. Tackling Data Bias in MUSIC-AVQA: Crafting a Balanced Dataset for Unbiased Question-Answering. arXiv:2310.06238.

Ma, C.-Y.; Lu, J.; Wu, Z.; AlRegib, G.; Kira, Z.; Socher, R.; and Xiong, C. 2019a. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint arXiv:1901.03035*.

Ma, C.-Y.; Wu, Z.; AlRegib, G.; Xiong, C.; and Kira, Z. 2019b. The regretful agent: Heuristic-aided navigation through progress estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6732–6740.

Majumdar, A.; Shrivastava, A.; Lee, S.; Anderson, P.; Parikh, D.; and Batra, D. 2020. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*, 259–274. Springer.

Nguyen, K.; Dey, D.; Brockett, C.; and Dolan, B. 2019a. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12527–12537.

Nguyen, K.; et al. 2019b. Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning. In *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 684–695. Hong Kong, China: Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; et al. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Pashevich, A.; et al. 2021. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15942–15952.

Paul, S.; Roy-Chowdhury, A. K.; and Cherian, A. 2022. AVLEN: Audio-Visual-Language Embodied Navigation in 3D Environments. *arXiv preprint arXiv:2210.07940*.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ram, A.; Prasad, R.; Khatri, C.; Venkatesh, A.; Gabriel, R.; Liu, Q.; Nunn, J.; Hedayatnia, B.; Cheng, M.; Nagar, A.; et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Shah, D.; Osinski, B.; Ichter, B.; and Levine, S. 2022. LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action. arXiv:2207.04429.

Siddhant, A.; and Lipton, Z. C. 2018. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. *arXiv preprint arXiv:1808.05697*.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2): 181–211.

Tan, H.; et al. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. *arXiv preprint arXiv:1904.04195*.

Thomason, J.; Murray, M.; Cakmak, M.; and Zettlemoyer, L. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, 394–406. PMLR.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, X.; Huang, Q.; Celikyilmaz, A.; Gao, J.; et al. 2019. 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*, 6629–6638.

Wijmans, E.; Kadian, A.; Morcos, A.; Lee, S.; Essa, I.; Parikh, D.; Savva, M.; and Batra, D. 2019. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*.

Xiao, Y.; and Wang, W. Y. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7322–7329.

You, C.; Chen, N.; Liu, F.; Ge, S.; Wu, X.; and Zou, Y. 2022. End-to-end Spoken Conversational Question Answering: Task, Dataset and Model. In *In Findings of NAACL 2022*.

Yu, Y.; Huang, W.; Sun, F.; Chen, C.; Wang, Y.; and Liu, X. 2022. Sound Adversarial Audio-Visual Navigation. *arXiv preprint arXiv:2202.10910*.

Zhou, G.; et al. 2023. NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. arXiv:2305.16986.

Zhu, W.; Hu, H.; Chen, J.; Deng, Z.; Jain, V.; Ie, E.; and Sha, F. 2020. Babywalk: Going farther in vision-and-language navigation by taking baby steps. *arXiv preprint arXiv:2005.04625*.

Zhu, Y.; Weng, Y.; Zhu, F.; Liang, X.; Ye, Q.; Lu, Y.; and Jiao, J. 2021. Self-Motivated Communication Agent for Real-World Vision-Dialog Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1594–1603.