AVLEN: Audio-Visual-Language Embodied Navigation in 3D Environments

Sudipta Paul[1]* Amit K. Roy-Chowdhury[1] Anoop Cherian[2]*

[1]University of California, Riverside (UCR) [2]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA

spaul007@ucr.edu cherian@merl.com

Video Computing Group, Center for Robotics & Intelligent Systems, UC Riverside Marlan and Rosemary Bourns College of Engineering

MITSUBISHI ELECTRIC — Changes for the Better

NEURAL INFORMATION PROCESSING SYSTEMS

Problem Statement

Given an agent

- that Is equipped with audio-visual sensors and is capable of embodied navigation in a realistic 3D environment (e.g., SoundSpaces), and
- which can interact with an oracle to **seek short navigation instructions,** where the instructions are provided in natural language;
- **the goal of agent** is to navigate and localize a sounding object in the scene in:
  (i) shortest number of steps and
  (ii) with the least number of instructions received from the oracle.

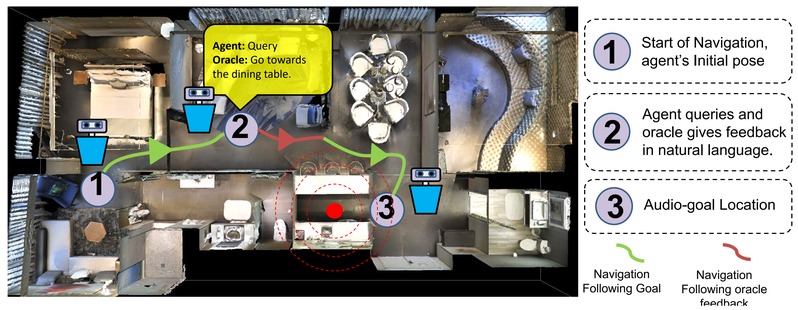


Figure 1: An illustration of our proposed AVLEN framework. The embodied agent starts navigating from location denoted ① guided by the audio-visual event at ③. At location ②, the learned policy for the agent decides to seek help from an oracle (e.g., because the audio stopped). The oracle provides a short natural language instruction for the agent to follow. The agent translates this instruction to produce a series of navigable steps to move towards the goal ③.

Our Contributions

- We introduce the **novel AVLEN task**, that unifies audio-visual embodied navigation with vision-and-language navigation towards building an audio-visual-language embodied navigation agent.
- We introduce a **novel multimodal hierarchical reinforcement learning** framework for solving the AVLEN task, that jointly learns three navigation policies:
  (i) a low-level policy to use audio-visual cues for navigation,
  (ii) a low-level policy to use vision-and-language instructions for navigation, and
  (iii) a high-level policy to decide which of (i) and (ii) to use given the scene, i.e., when to query the oracle.
- We provide experiments on the SoundSpaces dataset and show **state-of-the-art results,** especially when the audio-goal is intermittent, unheard, and under distractor sounds.

Prior Works

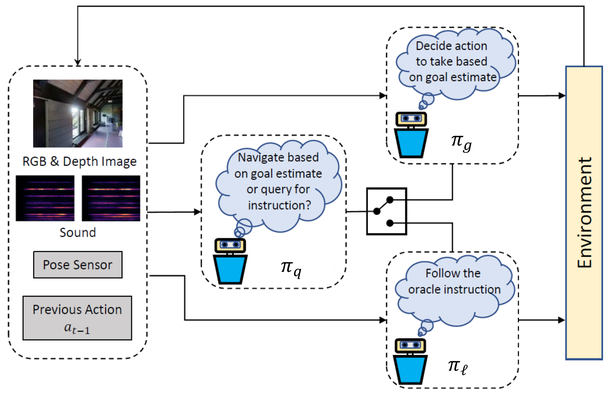
- **Just-ask, Chi et al., AAAI 2020** uses a threshold on the predictive navigation uncertainty for deciding when to ask for help from an oracle, while we learn to query in an end-to-end manner using a two-level hierarchy of policies.
- **Help Anna! Nhuyen and Daume' III, EMNLP 2019** assumes full observability of future actions of a selected policy to identify when the agent makes mistakes and then seek help, while AVLEN does not need such strong assumptions.
- **Cooperative Vision-and-Dialog Navigation, Thomason et al., CoRL 2019** focusses on dialogs for navigation but does not consider when to seek help from an oracle.

Task Formulation

We model the AVLEN task as a **Partially observable Markov decision process** (POMDP) characterized by $(S, A, T, R, O, P, \mathcal{V}, \gamma)$, where:

S: Set of agent states	O: Set of environment observations
A: Set of actions (right, left, forward, stop, query)	$P(o\|s', a)$: Probability of observing $\boldsymbol{o}$
$T(s'\|s, a)$: Transition probability	$\mathcal{V}$: language vocabulary
$R(s, a)$: Immediate reward for state-action pair	γ: Discount factor

Multimodal Hierarchical Deep Reinforcement Learning



Our goal is to learn a hierarchical policy π to maximize the value V, while minimizing the penalty ζ for querying the oracle, i.e.,

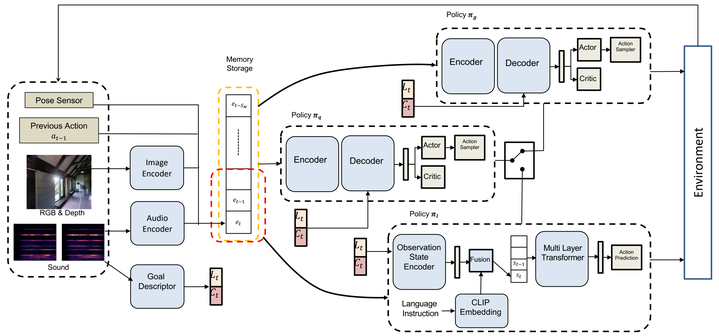
$$\arg\max_{\pi} V^{\pi}(b_0) \text{ where } V^{\pi}(b) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i \left(R'(b_{t+i}, a_{t+i}) - \zeta(t+i)\mathbb{I}(a_{t+i} = \text{query})\right) | b_t = b, \pi\right],$$

where,

$$b_{t+1}(s') = \eta P(o_{t+1}|s', a_t) \sum_{s \in \mathcal{S}} b_t(s) T(s'|s, a_t) \text{ and } R'(b, a) = \sum_{s \in \mathcal{S}} b(s) R(s, a)$$

The Bellman equation for using the hierarchical policies is given by:

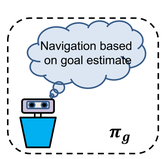
$$V^{\pi}(b) = \pi_q(\xi_g|b)\left[R'_g + \sum_{o' \in \mathcal{O}} P'(o'|b, \xi_g) V^{\pi}(b')\right] + \pi_q(\xi_\ell|b)\left[R'_\ell + \sum_{o' \in \mathcal{O}} P(o'|b, \xi_\ell) V^{\pi}(b')\right].$$



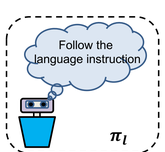
Detailed AVLEN multimodal hierarchical RL architecture

Generating Oracle Navigation Instructions: is produced using a Speaker model, Fried et al, NeurIPS, 2018, that takes shortest path actions to produce instruction words sequentially.

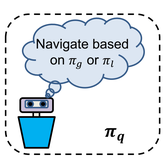
Policy Training AVLEN



❑Learning from reward signal
❑Rewards: distance to goal, reaching goal, time penalty



❑Learning from ground truth trajectory
❑Utilizes repurposed instruction-trajectory



❑No annotation or direct reward signals to train

Reward Design

$$\zeta_q(k) = \begin{cases} \frac{k \times (r_{neg} + \exp(-\nu))}{\nu} & k < K \\ r_{neg} + \exp(-k) & k \geq K, \end{cases} \quad \text{and} \quad \zeta_f(j) = \begin{cases} \frac{r_f}{j} & 0 < j < \tau \\ 0 & \text{otherwise}, \end{cases}$$

Experimental Analysis

We provide experiments on the benchmark *SoundSpaces* embodied navigation platform for the semantic audio-goal task in the setting where the audio is sporadic/intermittent. We consider three scenarios: (i) when the audio is from a category used during training (*heard)*, (ii) when audio is from a new category never heard during training (*unheard*), and (iii) *heard* but has multiple *distractor* sounds present in the scene.

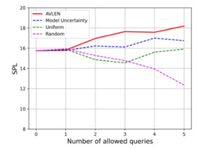
Table 1: Comparison of performances against state of the art in heard and unheard sound settings.

		Heard Sound					Unheard Sound				
	Feedback	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑
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. [14]	✗	29.3	23.7	**23.0**	11.3	14.4	15.9	12.3	11.6	12.7	8.0
Chen et al. [6]	✗	21.6	15.1	12.1	11.2	10.7	18.0	13.4	12.9	12.9	6.9
AV-WaN [5]	✗	20.9	16.8	16.2	10.3	8.3	17.2	13.2	12.7	11.0	6.9
SMT[12]+Audio	✗	22.0	16.8	16.0	12.4	8.7	16.7	11.9	10.0	12.1	8.5
SAVi [8]	✗	33.9	24.0	18.3	8.8	21.5	24.8	17.2	13.2	9.9	14.7
AVLEN	Language	**36.1**	**24.6**	19.7	**8.5**	**23.1**	**26.2**	**17.6**	**14.2**	**9.2**	**15.8**
AVLEN	GT Actions	**48.2**	**34.3**	**26.7**	**7.5**	**36.0**	**36.7**	**24.1**	**18.7**	**8.3**	**26.6**

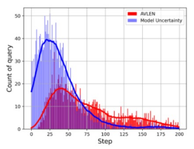
Figure 3: (a) Comparison of AVLEN performances against baselines and when-to-query approaches in the *presence of distractor sound*, (b) Performance (SPL) comparison against varying the number of allowed queries, and (c) Distribution of queries triggered against the time steps in episodes.

	Feedback	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑
Chen et al. [6]	✗	4.0	2.4	2.0	14.7	2.3
AV-WaN [5]	✗	3.0	2.0	1.8	14.0	1.6
SMT[12]+Audio	✗	4.2	2.9	2.1	14.9	2.8
SAVi [8]	✗	11.8	7.4	5.0	13.1	8.4
Random	Language	11.6	6.6	4.8	12.9	7.8
Uniform	Language	11.6	6.8	5.1	13.3	7.7
Model Uncertainty	Language	12.4	6.7	5.0	12.8	8.4
AVLEN	Language	**14.0**	**8.4**	**5.9**	**12.8**	**11.1**
AVLEN	GT Actions	**24.4**	**15.3**	**11.3**	**11.3**	**21.5**

(a) Comparisons under distractors



(b) Query sensitivity



(c) Query distribution

Ablation Analysis

We compare AVLEN when to query performance to:
i) Randomly decided,
ii) Uniformly spaced,
iii) Model uncertainty using π_g.

	$Step-1$	$Step-2$	$Step-3$
VLN-b (w/o instruction)	51.3	22.2	17.0
VLN-b	62.8	47.3	37.8
VLN-f	**65.9**	**55.5**	**45.3**

Table 3: Vision-language navigation performance.

Table 2: Comparisons in heard and unheard sound settings against varied query-triggering methods.

		Heard Sound					Unheard Sound				
	Feedback	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑
Random	Language	32.5	21.1	16.1	8.93	21.8	23.5	14.8	11.5	9.9	14.3
Uniform	Language	33.2	22.4	17.8	9.1	22.0	22.1	14.6	11.5	9.8	13.3
Model Uncertainty	Language	34.2	24.0	19.5	8.7	20.5	24.9	16.1	13.5	9.3	15.2
AVLEN	Language	**36.1**	**24.6**	**19.7**	**8.5**	**23.1**	**26.2**	**17.6**	**14.2**	**9.2**	**15.8**

Qualitative Results

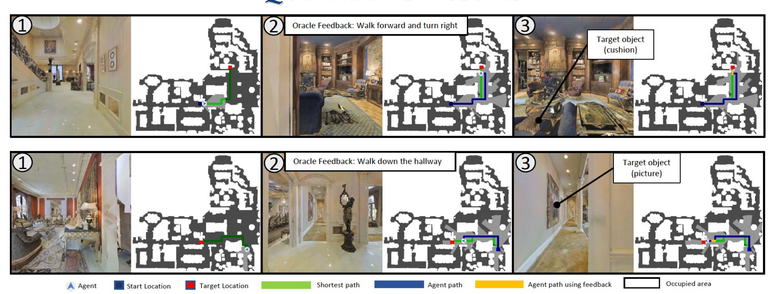


Figure 4: Two qualitative results from AVLEN's navigation trajectories. We show egocentric views and top down maps for three different viewpoints in agent's trajectory. The agent starts from ①, receives oracle help in ②, navigates to the goal in ③.

References

[1] Chi et al., Just ask: An interactive learning framework for vision and language navigation. AAAI, volume 34, pages 2459–2466, 2020.
[2] Khanh Nguyen and Hal Daum. III. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning, EMNLP, 2019
[3] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In Conference on Robot Learning, pages 394–406. PMLR, 2020.
[4] 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, NeurIPS, 2018

Acknowledgements

SP worked on this problem as part of a MERL internship. SP and AR are partially supported by the US Office of Naval Research grant N000141912264 and the UC Multi-Campus Research Program through award #A21-0101-S003.