Towards Audio-Visual Navigation in Noisy Environments: A Large-Scale Benchmark Dataset and an Architecture Considering Multiple Sound-Sources

Zhanbo Shi, Lin Zhang*, Linfei Li, Ying Shen

School of Computer Science and Technology, Tongji University
2111291@tongji.edu.cn, cslinzhang@tongji.edu.cn, cslinfeili@tongji.edu.cn, yingshen@tongji.edu.cn

Abstract

Audio-visual navigation has received considerable attention in recent years. However, the majority of related investigations have focused on single sound-source scenarios. Studies in this field for multiple sound-source scenarios remain underexplored due to the limitations of two aspects. First, the existing audio-visual navigation dataset only has limited audio samples, making it difficult to simulate diverse multiple sound-source environments. Second, existing navigation frameworks are mainly designed for single sound-source scenarios, thus their performance is severely reduced in multiple sound-source scenarios. In this work, we make an attempt to fill in these two research gaps to some extent. First, we establish a large-scale **BE**nchmark **D**ataset for **A**udio-**VI**sual **N**avigation, namely **BeDAViN**. This dataset consists of 2,258 audio samples with a total duration of 10.8 hours, which is more than 33 times longer than the existing audio dataset employed in the audio-visual navigation task. Second, we propose a new **E**mbodied **N**avigation framework for **MU**ltiple **S**ound-**S**ources **S**cenarios called **ENMuS^3**. There are mainly two essential components in ENMuS^3, the *sound event descriptor* and the *multi-scale scene memory transformer*. The former component equips the agent with the ability to extract spatial and semantic features of the target sound-source among multiple sound-sources, while the latter provides the ability to track the target object effectively in noisy environments. Experimental results on our BeDAViN show that ENMuS^3 strongly outperforms its counterparts with an order-of-magnitude improvement in success rates across diverse scenarios.

Code — https://github.com/ZhanboShiAI/ENMuS

Introduction

Embodied navigation, which requires an autonomous agent to solve challenging way-finding tasks by interacting with previously unseen environments, represents one of the most fundamental and essential components of embodied AI. In recent years, this technique has been employed in a wide range of applications, including but not limited to domestic service (Zhang, Zhang, and Shao 2021), warehousing (Gadd and Newman 2015), and logistics (Perdoch et al. 2015).

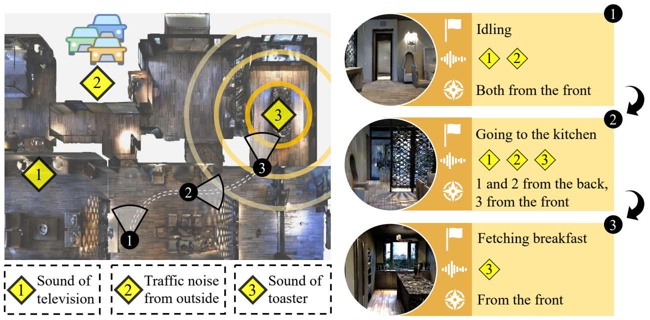

Figure 1: A typical scenario in real-world environments. In this context, the agent is instructed to fetch breakfast for its owner. This task requires the agent to be equipped with the capability to accurately discern the sound emitted by the toaster, despite the presence of other sound-sources such as television sound and traffic noise, and subsequently navigate to the kitchen relying on audio-visual observations.

To facilitate real-world applications, the current advances in embodied navigation tend to build an agent that utilizes egocentric vision to travel to the desired location (Anderson et al. 2018; Tang et al. 2022; Partsey et al. 2022), to search for a specific class of objects (Zhang et al. 2021; Majumdar et al. 2022; Qiao et al. 2023), or to cruise around the room following language instructions (Li and Bansal 2023; Gao et al. 2023). However, due to the complex structures of the indoor environments, the target objects or locations are frequently situated outside the field-of-view of the agent. Consequently, these visual-only approaches are inherently inefficient. In order to optimize the utilization of multiple sensory modalities for effective navigation, there has been an uptick in investigations exploring audio-visual navigation, which equips the agent with auditory capabilities and requires it to find a sounding object (Chen et al. 2020, 2021; Chen, Al-Halah, and Grauman 2021; Chen et al. 2023; Liu et al. 2024). In spite of the considerable efforts made in recent years, the deployment of these methods in multiple sound-source scenarios (multi-source scenarios in short) is hindered by limitations in two aspects.

- The existing audio dataset for audio-visual navigation (Chen, Al-Halah, and Grauman 2021) contains limited

[*] Corresponding Author

audio samples. This dataset comprises only one audio clip for each target sound category, making it difficult to simulate diverse multi-source scenarios. For example, in scenarios with multiple sound-sources of the same category, the same audio signals from different locations will be superimposed at the microphone, resulting in a false audio simulation.

- The majority of existing solutions for audio-visual navigation are designed for single sound-source scenarios (single-source scenarios in short) (Chen et al. 2020, 2021; Chen, Al-Halah, and Grauman 2021; Chen et al. 2023; Liu et al. 2024). It is notable, however, that real-world environments are commonly characterized by the presence of multiple sound-sources and background noise, as illustrated in Fig. 1. When deployed in such noisy scenarios, the performance of the existing methods significantly diminishes.

Taking the aforementioned limitations into consideration, in this paper, we establish a new dataset called **BeDAViN** (**BE**nchmark **D**ataset for **A**udio-**VI**sual **N**avigation) to facilitate the simulation of multiple sound-sources and propose an **E**mbodied **N**avigation framework for **MU**ltiple **S**ound-**S**ources **S**cenarios, namely **ENMuS^3**. Our contributions can be summarised as follows.

- To facilitate the study of audio-visual navigation in multi-source scenarios, we establish BeDAViN. BeDAViN consists of 2,258 audio samples encompassing 20 sound event categories and 4 noise categories, allowing the simulation of diverse multi-source scenarios. It is also worth noting that the total duration of this dataset is 10.8 hours, which is more than 33 times longer than the existing audio dataset (Chen, Al-Halah, and Grauman 2021) used for the audio-visual navigation tasks.
- A novel embodied navigation framework, ENMuS^3, is proposed in this paper to address the audio-visual navigation tasks in multi-source scenarios. ENMuS^3 incorporates two essential components, the *sound event descriptor* and the *multi-scale scene memory transformer*. The former component is designed to extract spatial and semantic information about the target sound-source from noisy binaural audio waveforms. The latter takes advantage of global interactions and local features of the scene memory across multiple resolutions to improve the navigation efficiency in noisy environments.
- Extensive experiments conducted on our BeDAViN demonstrate that our ENMuS^3 completely outperforms the state-of-the-art (SOTA) competitors in terms of the navigation success rate and the efficiency across diverse scenarios with different sound-source configurations.

Related Work

Vision-based Navigation. The general aim of embodied navigation is to identify a path from the starting position to a target position in 3D environments. To achieve this, the traditional navigation framework uses Simultaneous Localization and Mapping (SLAM) techniques to construct an occupancy grid as the agent wanders around the environment and then determines a path to the target with the wayfinding algorithms (Engel, Schöps, and Cremers 2014; Mur-Artal, Montiel, and Tardós 2015; Zhang et al. 2022). Unfortunately, these SLAM-based methods suffer from accumulated calculation errors along the mapping process, leading to poor performance in practical applications. With the advent of reinforcement learning (RL) techniques, recent advances tend to learn navigation policies directly from egocentric observations for a variety of purposes (Anderson et al. 2018; Zhang et al. 2021; Li and Bansal 2023; Gao et al. 2023). For instance, the PointGoal navigation task (Anderson et al. 2018) requires the agent to navigate to a specified position. Alternatively, in the ObjectGoal navigation task (Zhang et al. 2021), the agent is instructed to find the nearest instance of a specific object category. In recent years, numerous studies have been conducted on Vision-and-Language navigation (Li and Bansal 2023; Gao et al. 2023), in which the agent is expected to navigate to the destination following the natural language instructions. It should be noted that all the aforementioned RL-based methods restrict the agent with solely visual observations. As the target object is frequently situated outside the field-of-view of the agent, the navigation efficiency of these approaches is limited.

Audio-visual Navigation. Audio-visual navigation, also known as AudioGoal navigation (Chen et al. 2020), involves an agent navigating to a target object that emits sound by leveraging audio and visual observations in an unseen environment. To address this challenge, the majority of studies in recent years have attempted to train an end-to-end policy via RL techniques. For instance, AV-Nav (Chen et al. 2020) employs a multi-modal deep RL method to train the navigation policy that predicts low-level actions (e.g., moving forward, turning left, and turning right). To predict high-level waypoints, AV-Wan (Chen et al. 2021) is proposed to build both geometric and acoustic maps as the agent moves in the environments. Considering that the sounds in the above two studies are persistent, in order to locate and track sporadic target sounds, SAVi (Chen, Al-Halah, and Grauman 2021) incorporates a goal descriptor to remember the location of the target sound when it is inactive. For the same purpose, ORAN (Chen et al. 2023) is equipped with an omnidirectional information-gathering mechanism to collect visual-acoustic observations from different directions in unseen environments before decision-making. Additionally, inspired by methods of vision-and-language navigation, CAVEN (Liu et al. 2024) is proposed to perform audio-visual navigation with the help of the oracle. It is worth noting that, due to the limitations outlined regarding the dataset and frameworks for audio-visual navigation, the practical applications of these methods in real-world environments characterized by multiple sound-sources and background noise remain a significant challenge.

BeDAViN: A Benchmark Dataset for Audio-visual Navigation

To facilitate the study of audio-visual navigation, we establish a new large-scale benchmark dataset named BeDAViN.

Dataset	Total number of samples	Total duration of audio
SAVi-dataset (Chen, Al-Halah, and Grauman 2021)	144	1,157 seconds
BeDAViN (Ours)	2,258	10.8 hours

Table 1: The comparison of our BeDAViN and the existing dataset employed in audio-visual navigation.

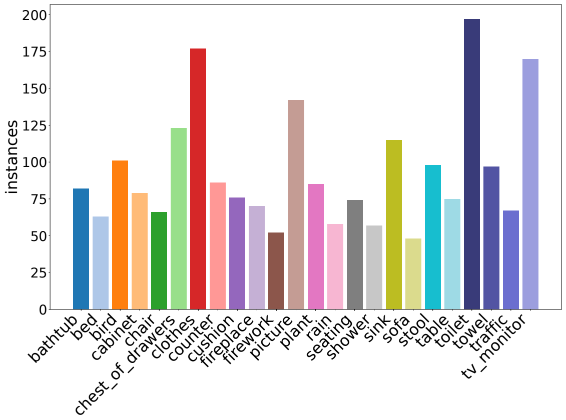

Figure 2: The number of audio samples in each category.

This dataset comprises 2,258 audio samples covering 24 sound event categories, allowing the simulation of diverse multi-source scenarios. A comparison of BeDAViN with the existing dataset for audio-visual navigation is summarized in Table 1, and a detailed distribution of audio samples in each category for our BeDAViN is illustrated in Fig. 2.

Categories Information. Our BeDAViN contains two sets of audio samples, one for target audio simulation and another for background noise generation. The target sound categories employed in the first set are similar to those used in the previous study of audio-visual navigation (Chen, Al-Halah, and Grauman 2021), with the only exception of the *gym equipment*. This category is omitted because its occurrence in simulation environments is insufficient to yield enough training and testing data. As for the background noise set, several categories that commonly serve as background noise in the field of sound event detection are adopted, including *rain* (raindrops), *bird* (bird vocalizations, bird calls, and bird song), *traffic noise* (roadway noise), and *fireworks*.

Sound Event Dataset Collection. To construct the sound event dataset comprising the aforementioned categories, a manual recording process was first conducted. The recording equipment utilized was a Tascam DR-40X with an X-Y microphone configuration and a built-in level of 70, as illustrated in Fig. 3 (a). In order to collect clear and noiseless samples, all audio files for the target sound categories were recorded within 1 meter of the sound-source in indoor environments. Fig. 3 (b) shows some instances of the recording process. As a result, 158 manually recorded 24-bit binaural wave files with a sampling rate of 96,000 Hz were obtained. To enrich this dataset, we subsequently selected audio clips referring to the manually recorded samples in two public datasets, AudioSet (Gemmeke et al. 2017) and FSD50K (Fonseca et al. 2022), which are widely used in the field of sound event detection. In the case of categories that have not been systematically collected in any datasets, such as *towel*, *cushion*, and *plant*, an exhaustive search was conducted in a public database, freesound.org, for audio clips of these categories under the Creative Commons licenses. To achieve this, the names and materials of these categories were used as keywords in order to facilitate the retrieval of relevant audio clips. Eventually, a large-scale sound event dataset was constructed including 2,258 audio clips with a total duration of 10.8 hours, which is 33 times longer than the existing audio dataset (Chen, Al-Halah, and Grauman 2021) employed in recent investigations for audio-visual navigation.

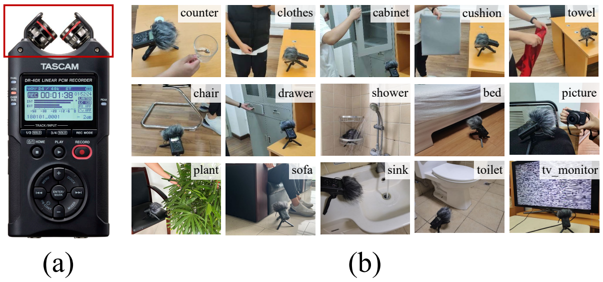

Figure 3: (a) The recording equipment Tascam DR-40X. The red rectangle indicates the X-Y microphone configuration. (b) Examples of the recording process. The sound event category of each instance is labeled in the top right corner of the image. All the recording processes for target categories were conducted within 1 meter of the target object in indoor environments.

Navigation Episodes Generation. To train the agent with our BeDAViN, 1.5 million navigation episodes were generated. Note that each episode comprises a set of parameters for simulating the navigation process in virtual environments, which can be defined by 1) the simulation scene, 2) the agent's start location and rotation, 3) the location and category of the target object, 4) the audio file name and the duration of the sound for target audio simulation. Two optional settings are also included in part of the episodes, 5) the audio file name and the duration of the sound for interfering sound simulation in multi-source scenarios, and 6) the noise file name for background noise generation in noisy scenarios. The detailed sound-source configurations for different scenarios are described in the experiments section. In an episode of a given scene, the agent's start location and the location of the target object were randomly selected with the guarantee that the geodesic distance between these two positions was greater than $4m$ and the ratio of Euclidean distance to geodesic distance was greater than 1.1. In addition, the audio files were also randomly sampled according to the sound category in our BeDAViN. Eventually, 1.5 million/3,000/3,000 episodes were collected for train/-

val/test splits in all 85 Matterport3D environments (Chang et al. 2017).

ENMuS^3: An Embodied Navigation Framework for Multi-Source Scenarios

Framework Overview

As illustrated in Fig. 4, to fulfill the audio-visual navigation task in multi-source scenarios, our ENMuS^3 first maps the local observations to an observation embedding for each time step using the Observation Encoder module. Specifically, to detect and locate the target sound-source among multiple sound-sources, this module includes a sound event descriptor, which equips the agent with the ability to extract both semantic and spatial features of the target sound-source. Subsequently, ENMuS^3 employs our multi-scale scene memory transformer to construct a multi-resolution memory representation. With the global and local features contained in this representation, ENMuS^3 can efficiently determine the next action of the agent in noisy environments.

Observation Encoder

The purpose of the Observation Encoder module is to transform the current scene observations into an observation embedding at each time step and then update the scene memory, which has been shown to be beneficial in long-horizon navigation tasks (Fang et al. 2019). The encoders employed in this module and their functions are described as follows.

Audio Encoder. The function of the audio encoder is to generate the low-level audio representation $\boldsymbol{e}_t^B$ with current audio observations for time step t. To achieve this goal, an audio feature extraction process is conducted as an initialization. First, the binaural waveforms are transformed into the left-channel and right-channel spectrogram, $\boldsymbol{S}^{\mathrm{L}}$ and $\boldsymbol{S}^{\mathrm{R}} \in \mathbb{C}^{T\times F}$, respectively, employing the short-time Fourier transform (STFT) with the Hamming window, where T is the number of time frames and F is the number of frequency bins configured in STFT. Then, the interaural spectrogram $\hat{\boldsymbol{S}}$ can be defined as follows,

$$\hat{\boldsymbol{S}} = \left[\hat{S}_{tf}, \cdots\right]_{T\times F}, \hat{S}_{tf} = S_{tf}^{\mathrm{L}}/S_{tf}^{\mathrm{R}}, \tag{1}$$

where S_{tf}^{L} and S_{tf}^{R} are the elements at the t-th time frame and the f-th frequency bin of $\boldsymbol{S}^{\mathrm{L}}$ and $\boldsymbol{S}^{\mathrm{R}}$, respectively. To mimic human auditory characteristics (Blauert 2013), the Interaural Phase Difference (IPD) $\boldsymbol{S}^{\mathrm{IPD}} \in \mathbb{R}^{T\times F}$ and Interaural Level Difference (ILD) $\boldsymbol{S}^{\mathrm{ILD}} \in \mathbb{R}^{T\times F}$ are derived from the interaural spectrogram as,

$$\begin{aligned}\boldsymbol{S}^{\mathrm{IPD}} &= \left[S_{tf}^{\mathrm{IPD}}, \cdots\right]_{T\times F}, S_{tf}^{\mathrm{IPD}} = \arg(\hat{S}_{tf})\\ \boldsymbol{S}^{\mathrm{ILD}} &= \left[S_{tf}^{\mathrm{ILD}}, \cdots\right]_{T\times F}, S_{tf}^{\mathrm{ILD}} = 20\log|\hat{S}_{tf}|,\end{aligned} \tag{2}$$

where $\arg(\cdot)$ and $|\cdot|$ denote the argument and the modulus of a complex number, respectively. Eventually, a four-channel acoustic feature $\boldsymbol{B} \in \mathbb{R}^{4\times T\times F}$ is obtained by combining the real parts of the left- and right-channel spectrograms, $\boldsymbol{S}_{real}^{\mathrm{L}}$ and $\boldsymbol{S}_{real}^{\mathrm{R}} \in \mathbb{R}^{T\times F}$, with IPDs and ILDs as,

$$\boldsymbol{B} = \mathrm{Concat}(\boldsymbol{S}_{real}^{\mathrm{L}}, \boldsymbol{S}_{real}^{\mathrm{R}}, \boldsymbol{S}^{\mathrm{IPD}}, \boldsymbol{S}^{\mathrm{ILD}}), \tag{3}$$

where $\mathrm{Concat}(\cdot)$ denotes the concatenation operation.

After obtaining this acoustic feature, a CRNN-based architecture (Adavanne et al. 2019) is employed taking it as the input to generate a low-level representation $\boldsymbol{e}_t^B \in \mathbb{R}^{N_b}$ of the audio observation for time step t, where N_b is the output dimension of the audio encoder.

Sound Event Descriptor. To facilitate the ability of the agent to distinguish the target sound-source among multiple sound-sources, we propose the sound event descriptor to extract the high-level audio representation $\boldsymbol{e}_t^D$. In detail, this module processes the output of the aforementioned audio encoder and generates a class-wise output, which contains the estimated categories of the active sound-sources in surroundings and their corresponding directions of arrival (DoAs), as shown in the red rectangle of Fig. 4. Specifically, the estimated DoAs are initially expressed in 3D coordinates, which can be represented as $\boldsymbol{x}_t$, $\boldsymbol{y}_t$, and $\boldsymbol{z}_t \in \mathbb{R}^{N_c}$ for time step t, where N_c is the number of sound event categories. Note that this estimation is expressed on the unit sphere centered at the agent, and the range of positions along each axis is $[-1, 1]$ relative to the origin of the agent coordinate system. For brevity, let $\mathcal{F}(x, y, z)$ denotes the following function,

$$\mathcal{F}(x, y, z) = \sqrt{x^2 + y^2 + z^2}. \tag{4}$$

Then the determination of whether the sound-source category $i \in [0, N_c)$ is active can be calculated as,

$$s_t^i = \begin{cases} 1, & \text{if} \quad \mathcal{F}(x_t^i, y_t^i, z_t^i) \geq \delta, \\ 0, & \qquad\qquad \text{otherwise}, \end{cases} \tag{5}$$

where δ is the threshold, which is set to 0.5 in the practice, x_t^i, y_t^i, and z_t^i are the i-th elements of $\boldsymbol{x}_t$, $\boldsymbol{y}_t$, and $\boldsymbol{z}_t$, respectively.

Considering that DoA estimation inevitably introduces localization errors, our sound event descriptor averages the DoAs over N_d time steps to minimize this error. To achieve this, these DoAs under the agent's local coordinate system are first transformed into the global coordinate system,

$$\begin{aligned}\gamma_t^i &= \arctan(y_t^i, x_t^i) + \mu\\ \theta_t^i &= \arcsin\left(\frac{z_t^i}{\mathcal{F}(x_t^i, y_t^i, z_t^i)}\right),\end{aligned} \tag{6}$$

where μ is the agent's current orientation in the global coordinate system, γ_t^i and θ_t^i represent the yaw and pitch of the i-th sound-source in the global coordinate system, respectively. After that, the yaw and pitch of the recent N_d time steps are fed into an LSTM network to minimize the location errors, and the hidden layers of this network are considered the high-level representation of the audio observations, denoted by $\boldsymbol{e}_t^D \in \mathbb{R}^{N_h}$, where N_h is the dimension of the processed hidden layers.

Visual/Pose/Action Encoders. We adopt a ResNet (He et al. 2016) as the visual encoder to generate the visual representation $\boldsymbol{e}_t^I \in \mathbb{R}^{N_i}$. To encode the current pose and the previous action of the agent, we employ two linear networks to produce the pose representation $\boldsymbol{e}_t^p \in \mathbb{R}^{N_p}$ and the action representation $\boldsymbol{e}_t^a \in \mathbb{R}^{N_a}$. Here, N_i, N_p, and N_a denote the output dimensions of their respective networks.

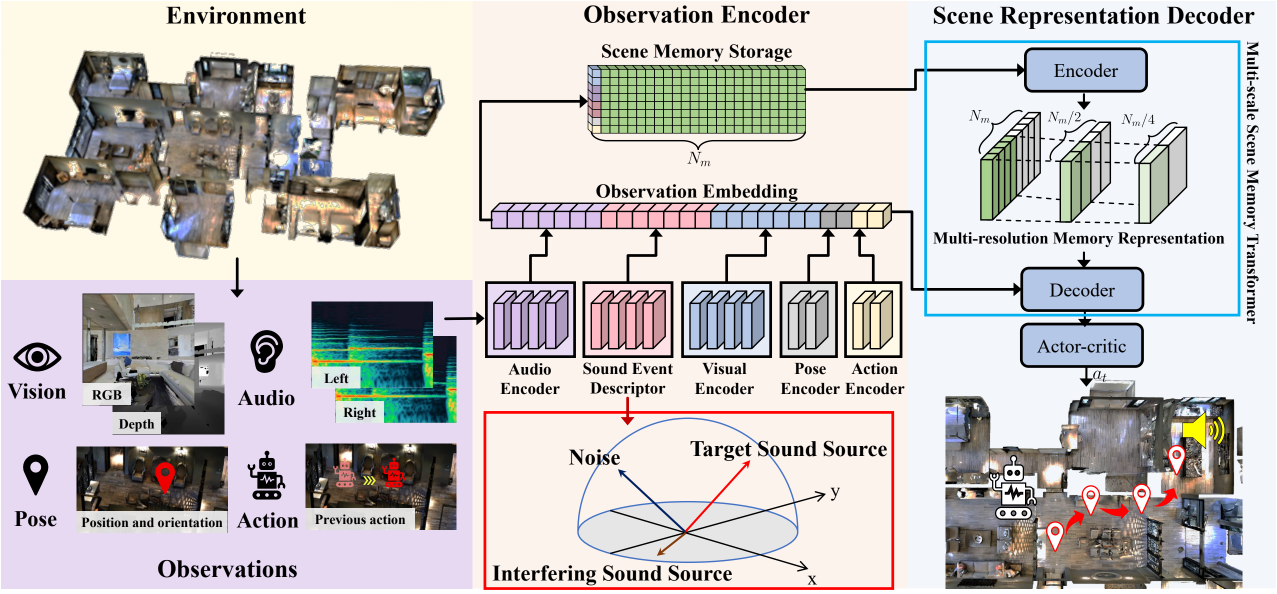

Figure 4: The framework overview of ENMuS^3. ENMuS^3 can be divided into two parts, an observation encoder to process multi-sensory scene observations and a scene representation decoder to predict the next action of the agent. Specifically, a sound event descriptor is proposed to accurately capture the semantic and spatial features of the target sound-source among multiple sound-sources, as shown in the red rectangle. In addition, to maintain tracking of the target sound-source in noisy environments, a multi-scale scene memory transformer is introduced, as illustrated in the blue rectangle.

Scene Memory Storage. To empower the agent with the capability to exploit the historical information to perform long-horizon navigation tasks more efficiently, we employ a scene memory storage to maintain the recent N_m scene observations. In detail, after obtaining all of the above-mentioned representations, the observation embedding $\boldsymbol{e}_t^O$ of the current time step t is constructed as,

$$\boldsymbol{e}_t^O = \text{Concat}(\boldsymbol{e}_t^B, \boldsymbol{e}_t^D, \boldsymbol{e}_t^I, \boldsymbol{e}_t^p, \boldsymbol{e}_t^a) \tag{7}$$

This observation embedding is then inserted into the memory storage $\boldsymbol{M}_t$. In order to achieve a balance between the navigation performance and the storage cost, $\boldsymbol{M}_t$ retains only the most recent N_m observation embeddings, which can be represented as,

$$\boldsymbol{M}_t = \{\boldsymbol{e}_i^O | i = \max\{0, t - N_m + 1\}, \ldots, t\}. \tag{8}$$

It should be noted that if the length of $\boldsymbol{M}_t$ exceeds the limitation N_m, the oldest embedding will be discarded upon the insertion of a new one.

Scene Representation Decoder

The scene representation decoder module decodes the current observation embedding and the scene memory storage with our multi-scale scene memory transformer and predicts the next action of the agent with an actor-critic network. As the target sound in the environment is sporadic, the agent must leverage both global interactions and local features to maintain tracking of the target sound-source in noisy environments when it is inactive. To achieve this, given the processed scene memory storage $\boldsymbol{e}_t^M$, our multi-scale scene memory transformer first generates a multi-resolution memory representation $\overline{\boldsymbol{e}_t^M}$ utilizing multiple convolution layers with the same kernel size and stride as follows,

$$\begin{aligned}
\boldsymbol{e}_t^{M^0} &= \boldsymbol{e}_t^M, \\
\boldsymbol{e}_t^{M^i} &= \text{Conv}_i\left(\boldsymbol{e}_t^{M^{i-1}}\right), \\
\overline{\boldsymbol{e}_t^M} &= \text{Concat}\left(\boldsymbol{e}_t^M, \boldsymbol{e}_t^{M^1}, \cdots, \boldsymbol{e}_t^{M^n}\right),
\end{aligned} \tag{9}$$

where $\boldsymbol{e}_t^{M^i}, i \in (0, n]$ is the scene memory representation generated by the i-th convolution layer and n is the number of convolution layers, as shown in the left column of Fig. 5. Next, to form the scene representation for the agent's action prediction, the generated $\overline{\boldsymbol{e}_t^M}$ is fed into the decoder of our multi-scale scene memory transformer in conjunction with the current scene observations. Specifically, this decoder employs two branches of processing in parallel, as illustrated in the right column of Fig. 5. The first one is a convolution branch which extracts the local audio and visual features from current scene observations. The second one utilizes the multi-head attention mechanism (Vaswani et al. 2017) to exploit the global interactions contained in the observation memory. The outputs of these two branches are then aggregated to constitute a state representation. After obtaining this state representation, an actor-critic network takes it as input to predict the action distribution and value for the current time step, and a categorical sampler samples the next action a_t from this action distribution. More training details of our ENMuS^3 are provided in the supplementary material.

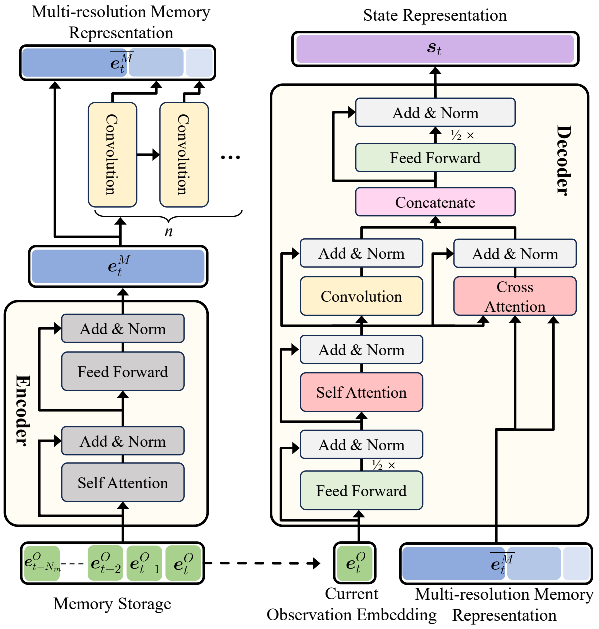

Figure 5: Overview of our multi-scale scene memory transformer. The left column illustrates the construction of the multi-resolution memory representation, while the right column details our decoder, which consists of a convolution branch and a cross-attention branch in parallel.

Experiments

Setup

Environments and Simulator. We adopted Matterport3D virtual indoor scenes (Chang et al. 2017), which are widely used in embodied navigation, to serve as the training and testing environments. As for the audio and visual simulator, we modified the SoundSpaces platform (Chen et al. 2020) by adding the generation pipelines for the interfering sound and the background noise.

Scenario Configurations. We conducted extensive experimentation on our BeDAViN in three types of scenarios, 1) single-source scenarios, where only the target object emitted sounds, 2) multi-source scenarios, in which multiple categories of sound events were active in the environment and a specific category of the sound event was considered the target sound-source, and 3) noisy scenarios, which were based on multi-source scenarios with additional everlasting background noise randomly selected from our BeDAViN. To ensure the generalizability of the experiments, all test results were averaged over 10 scenes from Matterport3D (Chang et al. 2017) with varying degrees of complexity, each comprising 100 episodes.

Evaluation Metrics. Similar to the existing studies (Chen et al. 2020; Chen, Al-Halah, and Grauman 2021; Chen et al. 2023), the following metrics were adopted to evaluate the performance of various audio-visual navigation schemes, 1) success rate (SR), 2) success weighted by inverse path length (SPL) (Anderson et al. 2018), 3) success weighted by inverse number of actions (SNA) (Chen et al. 2021), and 4) average distance to goal when episodes are finished (DTG).

Baseline Methods. We conducted a comparative analysis of our proposed ENMuS^3 against the following baselines.

- **Random**: A random policy uniformly samples one of three actions and increases the likelihood of executing *stop* as the number of actions increases.
- **Goal Follower**: A policy that first rotates the agent towards the direction of the target sound-source estimated by our sound event descriptor and then calls *move forward*.
- **ObjectGoal** (Batra et al. 2020): An end-to-end RL policy that takes RGB-D images and GPS compass as inputs to search the nearest instance of the target category. It is also given a one-hot encoding of the true category label as an additional input to search for the target object instance.
- **Av-Nav** (Chen et al. 2020): An end-to-end RL policy that encodes audio and visual observations with a GRU and predicts the low-level actions directly.
- **SAVi** (Chen, Al-Halah, and Grauman 2021): An end-to-end policy that adopts a ResNet to predict the location of the target sound-source and makes decisions utilizing a transformer-based architecture.
- **SMT** (Fang et al. 2019) **+ Audio**: A transformer-based policy that encodes the scene observations from the past N_m time steps into a memory representation and then predicts the agent's next action by decoding this representation. We modify this policy by also encoding the audio observations into its memory representation.

For a fair comparison, all the baseline methods and our ENMuS^3 used the same reward function and the same inputs if necessary, such as RGB-D images with a height and width of 128 pixels and the audio waveforms with a sampling rate of 16,000 Hz.

Quantitative Experimental Results

Table 2 presents the comparative navigation results of ENMuS^3 with SOTA audio-visual navigation methods on our BeDAViN. The optimal results are highlighted in bold. It is evident that our ENMuS^3 significantly outperforms all of its rivals. Specifically, ENMuS^3 achieves a notable increase in SR across all scenarios, surpassing the existing SOTA method by 13.1%, 7.1%, and 3.1%, respectively. This highlights the robust capability of our sound event descriptor in detecting and locating the target sound-source in diverse scenarios. Additionally, ENMuS^3 also demonstrates a considerable advancement in navigation efficiency, exhibiting a marked increase in both SPL and SNA. This suggests that our multi-scale scene memory transformer is capable of leveraging both global interactions and local features to identify shorter paths to the target object, thus improving the navigation efficiency. Further quantitative experiments (ablation study) are presented in the supplementary material.

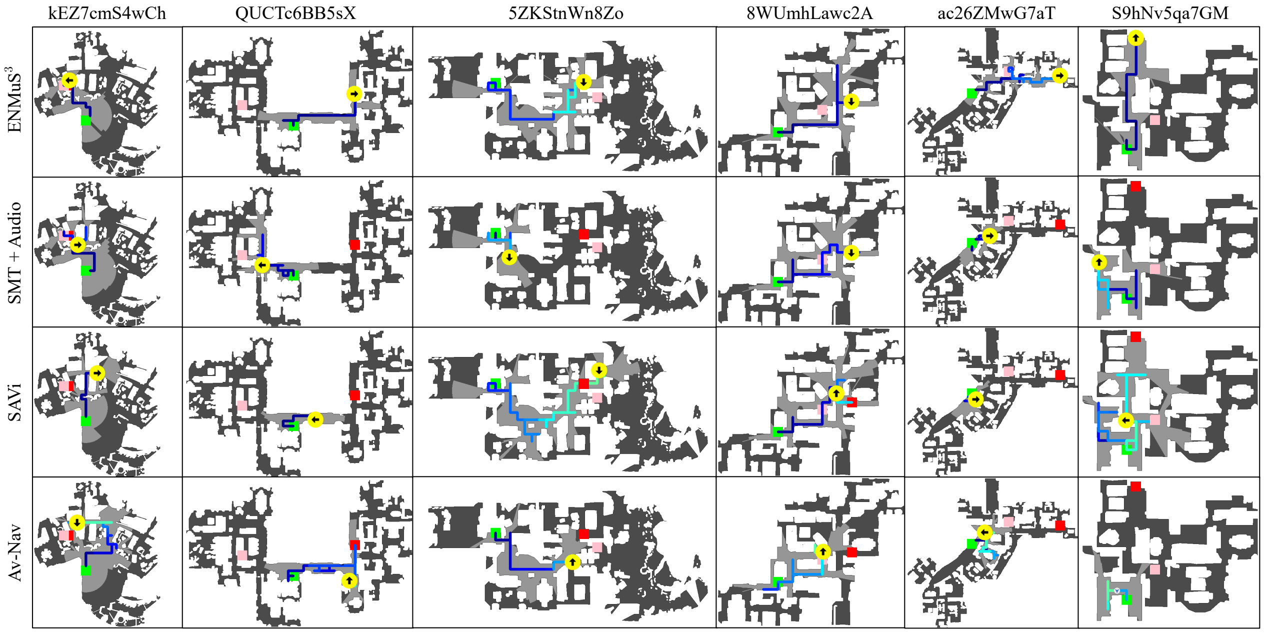

Figure 6: The navigation trajectories of our ENMuS^3 against other methods in multi-source scenarios. The green square, red square, and pink square indicate the start location of the agent, the location of the target object, and the location of the interfering sound-source, respectively. The yellow arrow shows the last location and orientation of the agent when it stops, and the blue line shows the navigation trajectories of the agent which gradually becomes lighter as the time step increases.

	Single-source Scenarios				Multi-source Scenarios				Noisy Scenarios			
	SR↑	SPL↑	SNA↑	DTG↓	SR↑	SPL↑	SNA↑	DTG↓	SR↑	SPL↑	SNA↑	DTG↓
Random	1.9	1.1	1.6	11.6	1.2	0.6	0.9	12.2	1.0	0.5	0.8	12.6
Goal Follower	2.1	1.1	1.7	11.5	1.2	0.7	1.0	12.0	1.3	1.0	1.2	12.5
ObjectGoal	2.8	1.0	1.4	10.7	3.4	1.2	1.7	11.0	2.2	0.6	1.0	11.1
Av-Nav	33.7	15.3	20.2	7.2	27.4	12.4	18.1	8.8	13.2	5.4	7.6	10.5
SAVi	26.3	11.1	14.4	7.3	19.9	8.5	9.3	8.9	12.0	6.4	7.1	10.4
SMT + Audio	66.3	31.9	47.3	3.9	37.9	17.7	24.5	6.9	14.1	6.1	9.4	10.1
ENMuS^3	**79.3**	**44.1**	**64.4**	**2.1**	**46.5**	**24.0**	**35.0**	**6.0**	**18.0**	**8.8**	**12.3**	**9.9**

Table 2: The performance of our ENMuS^3 and other compared methods on BeDAViN under various scenarios with different sound-source configurations.

Qualitative Experimental Results

Fig. 6 shows the navigation trajectories of our ENMuS^3 and other SOTA methods in multi-source scenarios. It can be observed that our ENMuS^3 is capable of completing the navigation tasks with more efficient paths. Especially, in *S9hNv5qa7GM* scene, our method reaches the target nearly following the shortest path, indicating the strong ability of our multi-scale scene memory transformer to track the target object in noisy environments. Furthermore, in situations where the target object is situated at a considerable distance from the agent's initial position, such as in *ac26ZMwG7aT* scene, our method can achieve the goal successfully with the help of our sound event descriptor, while other approaches get stuck within regions close to the starting point. More qualitative experiments in single-source scenarios and noisy scenarios are detailed in the supplementary material.

Conclusions

To facilitate audio-visual navigation in noisy environments, we introduced BeDAViN, a large-scale benchmark dataset containing 2,258 audio samples covering 24 sound event categories. Supported by BeDAViN, diverse scenarios with different sound-source configurations can be simulated, enabling the training and testing of the agent in multi-source environments. In addition, we proposed ENMuS^3, an embodied navigation framework for multi-source scenarios. ENMuS^3 is equipped with a sound event descriptor and a multi-scale scene memory transformer that significantly enhances the capacities of the agent to locate and track the target sound-source in challenging noisy environments. As existing methods of audio-visual navigation are developed in simulation environments, our future work will focus on deploying our ENMuS^3 in real-world applications.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62272343; in part by the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission under Grant 21SG23; and in part by the Fundamental Research Funds for the Central Universities.

References

Adavanne, S.; Politis, A.; Nikunen, J.; and Virtanen, T. 2019. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1): 34–48.

Anderson, P.; Chang, A.; Chaplot, D. S.; Dosovitskiy, A.; Gupta, S.; Koltun, V.; Kosecka, J.; Malik, J.; Mottaghi, R.; Savva, M.; and Zamir, A. R. 2018. On evaluation of embodied navigation agents. arXiv:1807.06757.

Batra, D.; Gokaslan, A.; Kembhavi, A.; Maksymets, O.; Mottaghi, R.; Savva, M.; Toshev, A.; and Wijmans, E. 2020. ObjectNav revisited: On evaluation of embodied agents navigating to objects. arXiv:2006.13171.

Blauert, J., ed. 2013. *Binaural Localization and Detection of Speakers in Complex Acoustic Scenes*. Springer.

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:1709.06158.

Chen, C.; Al-Halah, Z.; and Grauman, K. 2021. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15511–15520.

Chen, C.; Jain, U.; Schissler, C.; Gari, S. V. A.; Al-Halah, Z.; Ithapu, V. K.; Robinson, P.; and Grauman, K. 2020. SoundSpaces: Audio-visual navigation in 3D environments. In *Proceedings of the European Conference on Computer Vision*, 17–36.

Chen, C.; Majumder, S.; Al-Halah, Z.; Gao, R.; Ramakrishnan, S. K.; and Grauman, K. 2021. Learning to set waypoints for audio-visual navigation. arXiv:2008.09622.

Chen, J.; Wang, W.; Liu, S.; Li, H.; and Yang, Y. 2023. Omnidirectional information gathering for knowledge transfer-based audio-visual navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10993–11003.

Engel, J.; Schöps, T.; and Cremers, D. 2014. LSD-SLAM: Large-scale direct monocular SLAM. In *Proceedings of the European Conference on Computer Vision*, 834–849.

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.

Fonseca, E.; Favory, X.; Pons, J.; Font, F.; and Serra, X. 2022. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 829–852.

Gadd, M.; and Newman, P. 2015. A framework for infrastructure-free warehouse navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 3271–3278.

Gao, C.; Peng, X.; Yan, M.; Wang, H.; Yang, L.; Ren, H.; Li, H.; and Liu, S. 2023. Adaptive zone-aware hierarchical planner for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14911–14920.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 776–780.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Li, J.; and Bansal, M. 2023. Improving vision-and-language navigation by generating future-view image semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10803–10812.

Liu, X.; Paul, S.; Chatterjee, M.; and Cherian, A. 2024. CAVEN: An embodied conversational agent for efficient audio-visual navigation in noisy environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3765–3773.

Majumdar, A.; Aggarwal, G.; Devnani, B.; Hoffman, J.; and Batra, D. 2022. ZSON: Zero-shot object-goal navigation using multimodal goal embeddings. In *Proceedings of Advances in Neural Information Processing Systems*, 32340–32352.

Mur-Artal, R.; Montiel, J. M. M.; and Tardós, J. D. 2015. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5): 1147–1163.

Partsey, R.; Wijmans, E.; Yokoyama, N.; Dobosevych, O.; Batra, D.; and Maksymets, O. 2022. Is mapping necessary for realistic PointGoal navigation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17211–17220.

Perdoch, M.; Bradley, D. M.; Chang, J. K.; Herman, H.; Rander, P.; and Stentz, A. 2015. Leader tracking for a walking logistics robot. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, 2994–3001.

Qiao, Y.; Qi, Y.; Hong, Y.; Yu, Z.; Wang, P.; and Wu, Q. 2023. HOP+: History-enhanced and order-aware pre-training for vision-and-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8524–8537.

Tang, T.; Du, H.; Yu, X.; and Yang, Y. 2022. Monocular camera-based point-goal navigation by learning depth channel and cross-modality pyramid fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5422–5430.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, 5998–6008.

Zhang, S.; Song, X.; Bai, Y.; Li, W.; Chu, Y.; and Jiang, S. 2021. Hierarchical object-to-zone graph for object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15130–15140.

Zhang, T.; Zhang, L.; Chen, Y.; and Zhou, Y. 2022. CVIDS: A collaborative localization and dense mapping framework for multi-agent based visual-inertial SLAM. *IEEE Transactions on Image Processing*, 31: 6562–6576.

Zhang, Y.; Zhang, C.-H.; and Shao, X. 2021. User preference-aware navigation for mobile robot in domestic via defined virtual area. *Journal of Network and Computer Applications*, 173: 102885: 1–11.