Supplementary Material
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

Method Details

Problem Formulation in RL Framework

In the RL framework, the audio-visual navigation task can be formulated as a partially observable Markov decision process characterized by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{G}, O, T, R, \gamma)$, where $\mathcal{S}, \mathcal{A}, \mathcal{G}, R$, and γ represent the state space, action space, targets, reward function, and discount factor, $o_t = O(s_t)$ indicates the local observations at global state $s_t \in \mathcal{S}$ in decision step t, and $T(s_{t+1}|s_t, a_t)$ is the transition probability from s_t to s_{t+1} given action $a_t \in \mathcal{A}$. In this setup, ENMuS^3 adopts a policy $\pi(a|o, g)$ to produce its action a conditioned on its local observations o and target goal $g \in \mathcal{G}$.

Implementation Details

Our ENMuS^3 is implemented on a workstation with an AMD EPYC 7763 64-core Processor and an NVIDIA A100-SXM4-80GB GPU.

Training Details of ENMuS^3

Sound Event Descriptor Training. To train our sound event descriptor, we adopt a two-stage process. First, it is trained using wave files synthesized with the room impulse responses (RIRs) from the STARSS22 dataset (Politis et al. 2022) and audio samples from our BeDAViN. In this stage, we collect 900/300 wave files from 9 different rooms in STARSS22 for train/evaluation splits and train the sound event descriptor for 100 epochs to fully converge. As for the second stage, we finetune the sound event descriptor with audio samples simulated by Soundspaces (Chen et al. 2020) in Matterport3D environments (Chang et al. 2017). In the aforementioned two training stages, the SELD error (Shimada et al. 2022) is employed, which is widely used in the field of sound event localization and detection. More parameters utilized for sound event descriptor training are listed in Table 1. Note that the finetuned sound event descriptor will be frozen during the policy training.

Other Encoders Training. Similar to the previous work (Chen, Al-Halah, and Grauman 2021), we train the audio encoder, visual encoder, pose encoder, and action encoder with the size of the scene memory storage $N_m = 1$, and freeze them during the policy training.

*Corresponding Author

Parameter	Value	Parameter	Value
Sampling rate	16,000 Hz	Label sequence length	50
Hop length	160	Feature sequence length	500
Label hop length	1600	Batch size	256
Window length	160	Dropout rate	0.05
Number of FFT	400	Learning rate	2.5×10^{-4}
Max audio length	60 s	DoA threshold	20°
Number of Mel bins	64	Number of classes	20

Table 1: The parameters employed in the training stage of the sound event descriptor.

Policy Training. We use the decentralized distributed proximal policy optimization (DD-PPO) (Wijmans et al. 2020) to train the navigation policy with the full memory size $N_m = 160$. A value loss for the state-value prediction and a cross-entropy loss for the action distribution are adopted in this stage. We train the policy with a learning rate of 2.5×10^{-4} for 200 million steps to fully converge.

More Experiments

Ablation Study

Ablation of Sound Event Descriptor. Table 2 shows the ablation experimental results of our sound event descriptor on BeDAViN. It is obvious that in all cases there is a notable decline in the success rate (SR) of the part in the absence of the sound event descriptor. Especially, in the multi-source scenarios where the capability to discern the target sound-source among multiple sources is of paramount importance, the SR is observed to diminish considerably by 4.2%.

Ablation of Multi-scale Scene Memory Transformer. Next, we conducted experiments by ablating the multi-scale scene memory transformer. In order to clearly show the effect of this module on navigation efficiency, the number of actions taken by the agent to finish the navigation process (NA) was also evaluated. It can derived from Table 2 that this

	Single-source Scenarios					Multi-source Scenarios					Noisy Scenarios				
	SR↑	**SPL↑**	**NA↓**	**SNA↑**	**DTG↓**	**SR↑**	**SPL↑**	**NA↓**	**SNA↑**	**DTG↓**	**SR↑**	**SPL↑**	**NA↓**	**SNA↑**	**DTG↓**
w/o SEDesc	77.7	45.6	114.9	62.8	2.6	42.3	22.9	131.2	30.8	6.4	17.1	**9.0**	120.0	**13.6**	10.1
w/o MSMT	**81.1**	**46.2**	110.5	**67.0**	**1.9**	**47.6**	**25.3**	124.1	34.9	**5.6**	17.8	8.0	138.1	12.3	**9.4**
ENMuS^3	79.3	44.1	**100.2**	64.4	2.1	46.5	24.0	**116.9**	**35.0**	6.0	**18.0**	8.8	**105.8**	12.3	9.9

Table 2: Ablation experimental results of our ENMuS^3 on BeDAViN. SEDesc represents the *sound event descriptor*, and MSMT stands for the *multi-scale scene memory transformer*.

	ER↓	**F1↑**	**DE↓**	**FR↑**
Single-source	59.1	45.5	15.7	47.4
Multi-source	83.6	24.7	24.9	29.0
Noisy	95.4	16.0	63.2	20.9

Table 3: Evaluation of the confidence of our sound event descriptor in different scenarios.

	SR↑	**SPL↑**	**SNA↑**	**DTG↓**
Random	1.4	3.5	1.2	17.0
Goal Follower	1.5	0.8	0.6	16.7
ObjectGoal	2.1	1.1	1.7	11.5
Av-Nav	18.0	13.4	12.9	12.9
SAVi	24.8	17.2	13.2	9.9
SMT + Audio	16.7	11.9	10.0	12.1
ENMuS^3	**26.8**	**18.8**	**16.1**	**9.8**

Table 4: Navigation performance on SAVi-dataset in single-source scenarios without background noise.

module can substantially reduce the NA while maintaining a comparable SR. That is to say, our multi-scale trajectory memory transformer can exploit both global and local features to accomplish navigation tasks more efficiently.

Analysis of Sound Event Descriptor Accuracy. To elucidate the different effects of our sound event descriptor on the navigation performance in different scenarios, a further investigation of its accuracy in sound event detection (SED) and sound-source localization (SSL) was performed with four additional metrics. In detail, the error rate (ER) and F1-score (F1) calculated in segments of one second with no overlap proposed in (Virtanen, Plumbley, and Ellis 2018) were adopted for SED. As for SSL, two frame-wise metrics were applied. The first one is a conventional directional error (DE) expressing the angular distance between the reference and predicted directions of arrival of the sound-sources. The second one is the frame recall (FR) presented in (Adavanne et al. 2019) to account for the time frame where the number of estimated and reference sound-sources are unequal.

Table 3 shows the accuracy of our sound event descriptor in single-source, multi-source, and noisy scenarios. As can be observed from this table, in both single- and multi-source scenarios, the DEs are relatively small within 30 degrees. It is anticipated that these errors will be less significant in practice, given that our sound event descriptor will average the estimated DoAs over N_d time steps. In addition, although

	SR↑	**SPL↑**	**SNA↑**	**DTG↓**
Random	0.8	0.5	0.3	18.0
Goal Follower	1.5	0.8	0.6	16.7
ObjectGoal	1.8	1.2	1.2	15.5
Av-Nav	4.0	2.4	2.0	14.7
SAVi	11.8	7.4	5.0	13.1
SMT + Audio	4.2	2.9	2.1	14.9
ENMuS^3	**12.6**	**8.0**	**6.8**	**12.5**

Table 5: Navigation performance on SAVi-dataset in single-source scenarios with background noise.

the ERs of SED initially appear to be high, we find that this is primarily caused by the low performance in several sound-similar categories, e.g., *bed*, *chair*, and *setting*, which emit sounds of wood squeaking. That is to say, in both single- and multi-source scenarios, our sound event descriptor is capable of accurately estimating the semantic (sound event category) and spatial (direction of arrival) features of the target sound-source, thus improving the success rate of audio-visual navigation. In noisy scenarios, our sound event descriptor has a diminished impact on navigation, mainly due to the relatively large errors in SED and SSL.

Additional Navigation Results on SAVi-dataset. To further evaluate the navigation ability of our ENMuS^3 in single-source scenarios, we conduct extensive experiments on SAVi-dataset (Chen, Al-Halah, and Grauman 2021) in two types of scenarios, 1) single-source scenarios without background noise, and 2) single-source scenarios with background noise (referred to as the distractor sound-source in this dataset). Table 4 and Table 5 show the performance of our ENMuS^3 against other methods in these scenarios, respectively. As can be observed from these tables, our ENMuS^3 outperforms its nearest competitor, SAVi (Chen, Al-Halah, and Grauman 2021), with a 2.0% and 0.8% absolute gain on SR in two scenarios, respectively, which demonstrates the strong ability of ENMuS^3 to perform audio-visual navigation tasks in single-source scenarios.

Additional Navigation Trajectories on BeDAViN. In this section, we present additional navigation trajectories in single-source scenarios (Fig. 1) and noisy scenarios (Fig. 2) on our BeDAViN. It is clear that our ENMuS^3 is capable of finding a way to the target object with more efficient paths and dealing with the challenging navigation tasks where the target object is situated at a considerable distance from the agent's start position in both of these two scenarios.

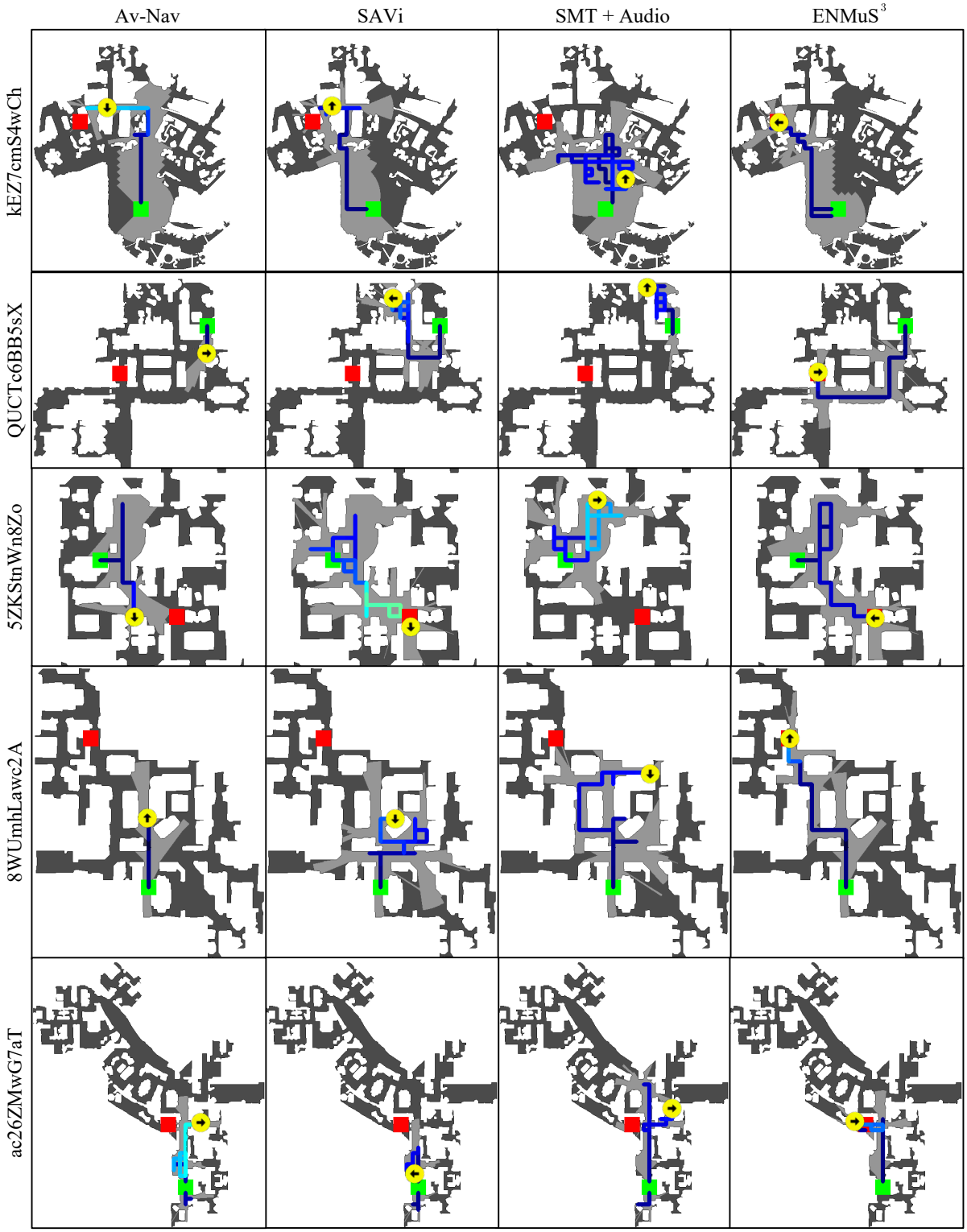

Figure 1: The navigation trajectories of our ENMuS^3 against the SOTA in single-source scenarios. The green square and red square indicate the start location of the agent and the target object location, respectively. The yellow arrow shows the last location and orientation of the agent when it stops, and the blue line shows the agent's navigation trajectories which gradually becomes lighter as the time step increases. It can be observed that our ENMuS^3 can reach the target with a shorter path as well as deal with challenging tasks where the target is situated at a considerable distance from the agent's initial positions.

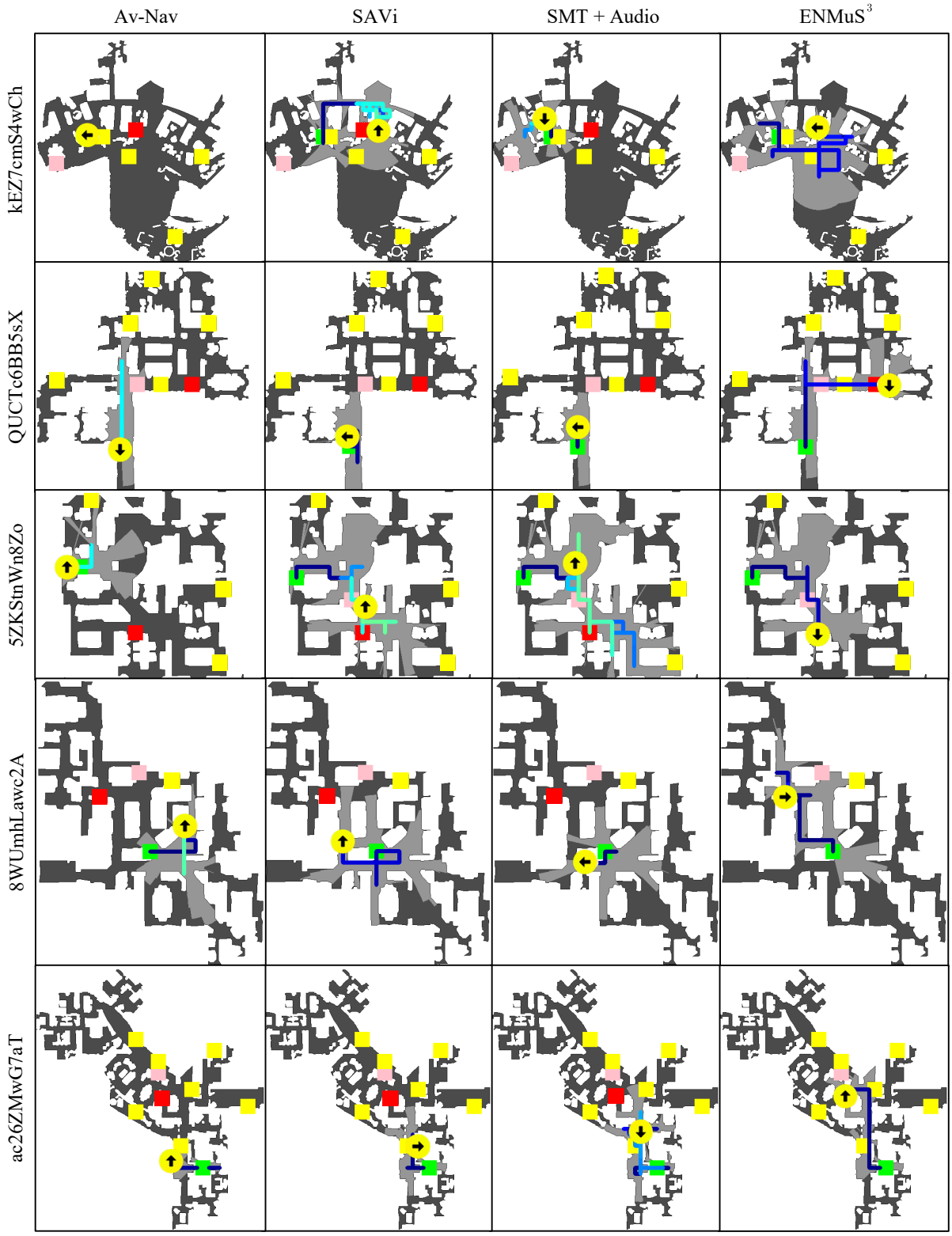

Figure 2: The navigation trajectories of our ENMuS^3 against the SOTA in noisy scenarios. The additional yellow squares indicate the locations of the background noise. It is obvious that our ENMuS^3 demonstrates a robust navigation ability in noisy scenarios.

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.

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.

Politis, A.; Shimada, K.; Sudarsanam, P.; Adavanne, S.; Krause, D.; Koyama, Y.; Takahashi, N.; Takahashi, S.; Mitsufuji, Y.; and Virtanen, T. 2022. STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. arXiv:2206.01948.

Shimada, K.; Koyama, Y.; Takahashi, S.; Takahashi, N.; Tsunoo, E.; and Mitsufuji, Y. 2022. Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 316–320.

Virtanen, T.; Plumbley, M. D.; and Ellis, D., eds. 2018. *Datasets and evaluation*. Springer.

Wijmans, E.; Kadian, A.; Morcos, A.; Lee, S.; Essa, I.; Parikh, D.; Savva, M.; and Batra, D. 2020. DD-PPO: Learning near-perfect pointGoal navigators from 2.5 billion frames. arXiv:1911.00357.