

Semantic Audio-Visual Navigation in Continuous Environments

Supplementary Material

1. Video Demonstrations

Video demonstrations with binaural audio recorded on the SAVN-CE dataset are provided to intuitively illustrate the task setup and the navigation performance of MAGNet. For better perception of spatial sound, headphones are recommended. Each video is named using the format: *VideoID_SceneID_EpisodeID_GoalCategory_SPL.mp4*.

ID 0 is from AV-Nav, while IDs 1-3 and 4-6 are recordings of MAGNet (ours) under *Clean Environments* and *Distracted Environments*, respectively. A brief summary of each video is provided below:

- ID 0: The agent freely explores the environment at the beginning, then navigates toward the goal during the sound-emitting period, but loses track once the sound ceases.
- ID 1: The agent successfully reaches the goal during the sound-emitting period.
- ID 2: The agent successfully reaches the goal after the sound ceases.
- ID 3: The agent fails to reach the goal after the sound ceases due to stopping prematurely.
- ID 4: The agent successfully reaches the goal.
- ID 5: The agent incorrectly stops near the distractor.
- ID 6: The agent fails to stop near any sound source, neither the goal nor the distractor.

2. Implementation Details

Experiments are conducted on two Linux servers equipped with dual Intel(R) Xeon(R) Platinum 8378A CPUs (64 cores, 128 threads in total) and eight NVIDIA A800-SXM4-40GB GPUs, among which four are utilized in our experiments. We employ DD-PPO implemented in PyTorch for distributed training, with each learner interacting with 10 parallel environments. The simulator operates at approximately 200 frames per second (fps) under *Clean Environments* and 120 fps under *Distracted Environments*. In total, completing all training and evaluation processes (including ours and the baselines) takes more than two months.

3. Audio Simulation Details

Our simulator is developed based on the SoundSpaces 2.0 platform [6]. To compute binaural audio, we convolve the single-channel source sound with the room impulse responses (RIRs) of the current time step, while also accumulating residual contributions from all previous time steps up to the present.

At time step t , let the room impulse response (RIR) \mathbf{h}_t have a length of L_t and the source sound s_t have a length

of M (here $M = 4000$). For binaural audio rendering, we denote $\mathbf{h}_{l,t}$ and $\mathbf{h}_{r,t}$ as the left and right channel RIRs, respectively (we omit the channel index below for clarity). The convolved waveform \mathbf{x}_t is given by:

$$\mathbf{x}_t[n] = s_t[n] * \mathbf{h}_t[n] = \sum_{k=0}^{L_t-1} s_t[k] \mathbf{h}_t[n-k], \quad (1)$$

where $*$ denotes convolution, and $n = 0, \dots, L_t + M - 1$ is the index of the convolved waveform samples.

Since the convolved waveform not only affects the current step but also propagates into subsequent steps, the audio observation \mathbf{y}_t is computed by accumulating the residual signals from all past steps:

$$\mathbf{y}_t[n] = \sum_{\tau=0}^t \mathbf{x}_\tau[(t-\tau)M + n], \quad (2)$$

where $n = 0, \dots, M-1$ indexes the audio observation samples, and \mathbf{x}_τ is zero-padded when the index is out of range.

If a distractor is present, an additional audio sensor is instantiated at its location to capture its signal, and its contribution is added as:

$$\mathbf{y}_t[n] = \mathbf{y}_{t,g}[n] + \mathbf{y}_{t,d}[n], \quad (3)$$

where $\mathbf{y}_{t,g}[n]$ and $\mathbf{y}_{t,d}[n]$ are the audio observations from the goal and the distractor, respectively.

4. Acoustic Feature Extraction

Given the audio observation $\mathbf{y}[n]$, the complex spectrogram is computed using short-time Fourier transform (STFT):

$$\mathbf{Y}(t, f) = \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{y}[k + tH] e^{-\frac{j2\pi f k}{K}}, \quad (4)$$

where $k = 0, \dots, K-1$ is the sample index, $K = 512$ is the STFT window length, $H = 160$ is the hop length, t is the frame index, and f is the frequency bin index. For simplicity, we drop the frame index and frequency bin index in the following equations. The mean magnitude spectrogram, the inter-channel phase difference (IPD), and the inter-channel level difference (ILD) are computed as:

$$\begin{aligned} \mathbf{Y}_{\text{mean}} &= (|\mathbf{Y}_l| + |\mathbf{Y}_r|) / 2, \\ \mathbf{Y}_{\text{IPD}} &= \arg[\mathbf{Y}_l] - \arg[\mathbf{Y}_r], \\ \mathbf{Y}_{\text{ILD}} &= |\mathbf{Y}_l| / |\mathbf{Y}_r|, \end{aligned} \quad (5)$$

where $\arg[\cdot]$ denotes the phase angle of a complex number, and l and r indicate the left and right channels, respectively. The acoustic feature \mathbf{A} is defined as:

$$\mathbf{A} = [\mathbf{Y}_{\text{mean}}, \sin(\mathbf{Y}_{\text{IPD}}), \cos(\mathbf{Y}_{\text{IPD}}), \log_{10}(\mathbf{Y}_{\text{ILD}})], \quad (6)$$

where \sin , \cos , and \log_{10} denote the sine, cosine, and base-10 logarithm functions, respectively. These transformations are applied to avoid angle wraparound and normalize the values. In the frequency domain, we use frequency bins up to the Nyquist frequency while excluding the DC bin. Thus, the acoustic feature \mathbf{A} has a shape of $4 \times 256 \times 26$.

5. Training Details of the Baselines

Random. We estimate the action distribution by analyzing oracle actions on the train split, obtaining 71% *MoveForward*, 14% *TurnLeft*, 14% *TurnRight*, and 1% *Stop*. During testing, actions are sampled from this distribution [8].

ObjectGoal. We adopt the model proposed in AV-Nav, removing the audio encoder and augmenting the input with a one-hot encoding of the goal category.

AV-Nav. We use the official implementation of AV-Nav [3]. To align the audio encoder input size, we only downsample the audio spectrogram along the frequency axis by a factor of 4 (rather than along both frequency and time). This adjustment is necessary because the step time in our setup is 0.25 s rather than 1 s. The next two baselines follow the same procedure.

SMT + Audio. We adopt the model proposed in SAVi [4], removing the goal descriptor network.

SAVi. We follow the architecture and training procedure in the official implementation of SAVi [4]. For pretraining the category classifier, we generate room impulse responses (RIRs) in Matterport3D scenes [2], using 59/11/15 scenes for train/val/test splits. For each scene, 1,000 RIRs are generated with source-receiver distances between 2 and 10 m. Goal sounds from the 21 semantic categories are convolved with the generated RIRs and clipped to match the simulation step time (0.25 s, 4000 samples), yielding 5.8M/1.1M/1.6M samples for train/val/test. The classifier is trained for 50 epochs with a learning rate of 1×10^{-3} , a batch size of 1024, and cross-entropy loss. In distractor cases, random additional RIRs and distractor sounds from the 102 periodic sounds in SoundSpaces [3] are sampled and mixed with goal sounds. The best model is selected based on validation performance, achieving a precision of 0.94 when only the goal sound is present and 0.62 when both goal and distractor sounds are present on the test split.

Excluded Methods. The following methods are not included in our experiments:

- **AV-WaN [5]:** AV-WaN dynamically sets waypoints and learns end-to-end within the navigation policy, and it is equipped with an acoustic memory that provides a structured and spatially grounded record of what the agent has heard as it moves. We encounter difficulties reproducing the method in the grid-based discrete setting, and adapting it to continuous environments proves challenging due to its tight coupling with grid points. According to the navigation results reported in SAVi [4], AV-WaN shows inferior performance compared with SAVi. Since our proposed MAGNet already surpasses SAVi in our continuous setting, we do not attempt further adaptation of AV-WaN.
- **ENMuS³ [12]:** This method is most similar to our proposed MAGNet. It employs: 1) a sound event descriptor that predicts the goal’s position and category in the ACCDOA format [13], and 2) a multi-scale scene memory transformer that decodes the current observation embedding and the scene memory to predict the next action. The descriptor is first trained using STARSS22 [11] and their proposed BeDAViN dataset, then fine-tuned with audio simulated by Soundspaces [3], after which it remains frozen during policy training. Because the pretraining and fine-tuning codes are not released, we are unable to retrain the descriptor. We attempt to modify the released code to fit our setup, but encounter issues such as mismatches between input and output dimensions. Even after adjusting the code to address these differences and initializing the descriptor with the provided checkpoint, either frozen or trained during policy learning, the performance remains unsatisfactory. Considering the substantial adaptation challenges, and that the reported navigation results of ENMuS³ on the SAVi dataset are not significantly higher than SAVi, we do not pursue further attempts to include ENMuS³ as a baseline.

6. Additional Navigation Trajectories

To gain insight into the observed differences in navigation performance, we visualize additional trajectories of our method and the baselines. Fig. 1 and Fig. 2 show the navigation trajectories under *Clean* and *Distracted Environments*, respectively.

A closer analysis of the navigation trajectories reveals three main causes of failure for our method: 1) the agent fails to locate the goal in complex continuous environments; 2) the agent continues navigating toward the goal after the sound ceases but stops prematurely, failing to reach within 1 m of the goal (the 1 m success criterion may be too strict in large environments); and 3) the agent fails to distinguish the goal from the distractor or to resist the influence of the distractor, often being misled by its stronger or more salient acoustic cues.

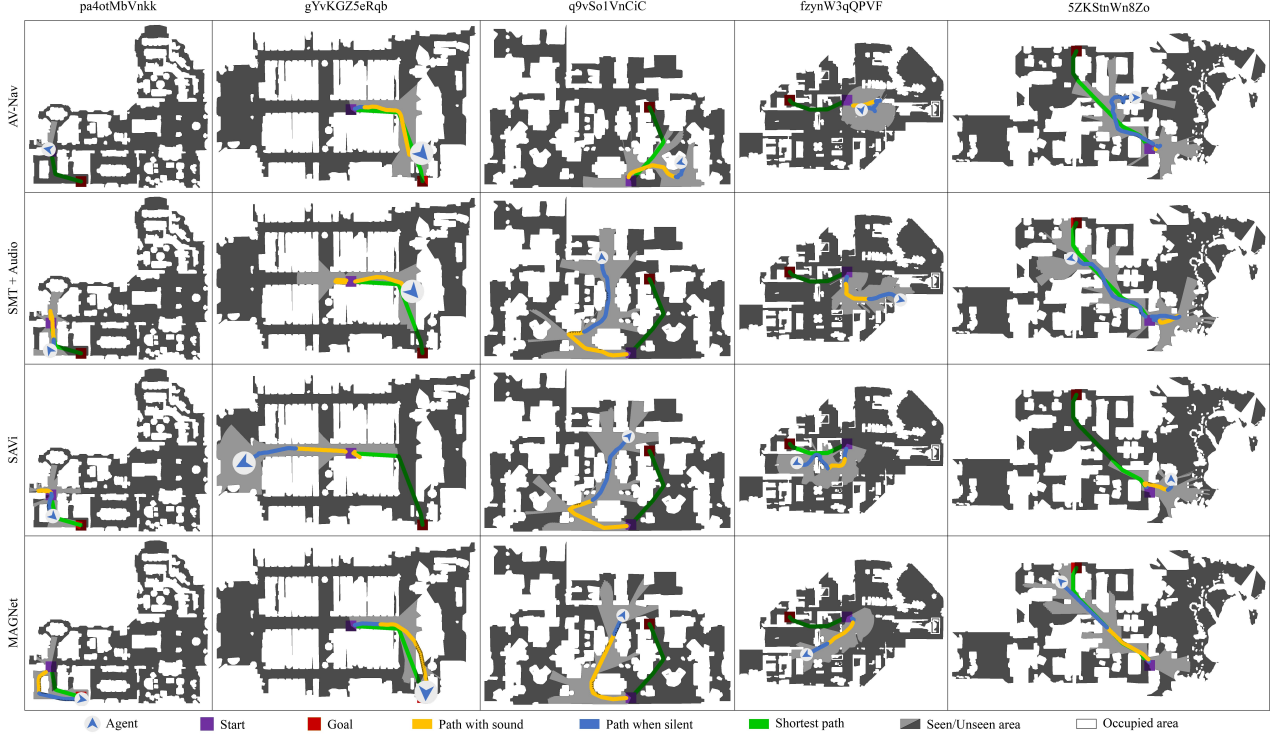


Figure 1. Navigation trajectories of our method and the baselines under *Clean Environments*.

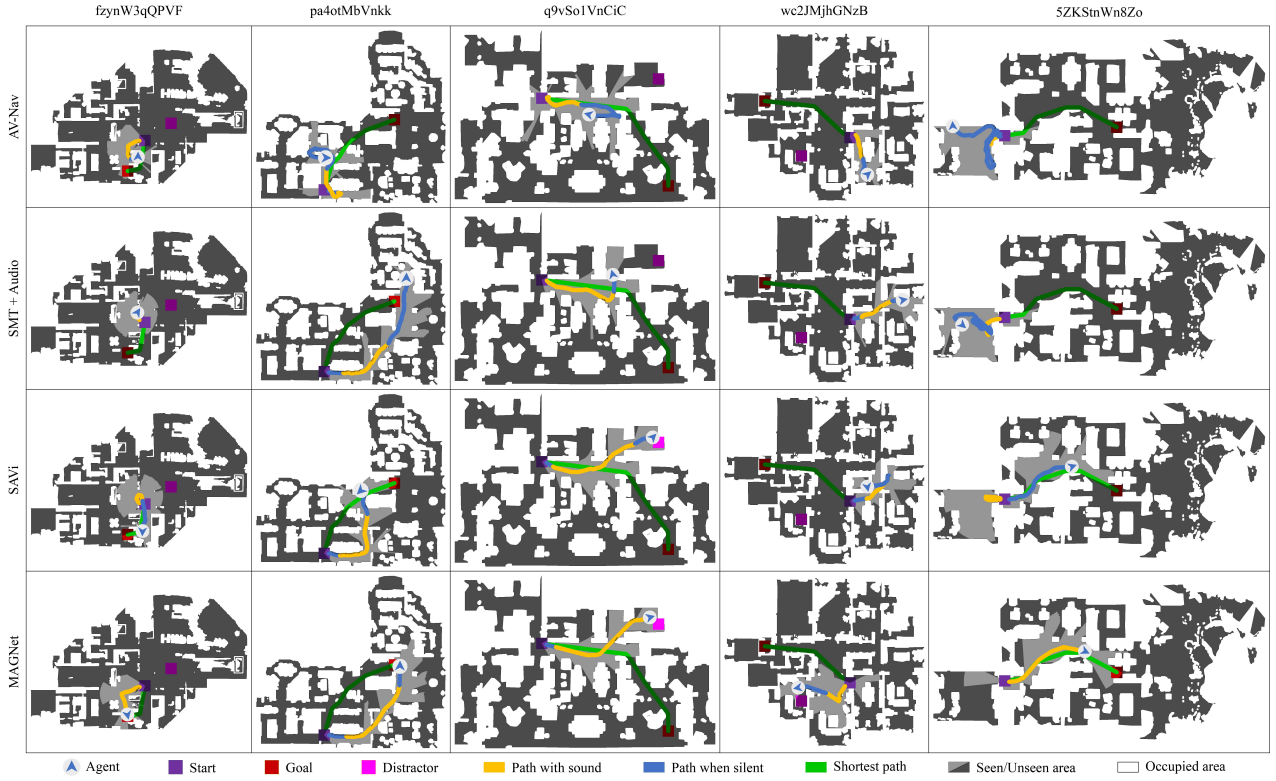


Figure 2. Navigation trajectories of our method and the baselines under *Distracted Environments*. Note that the goal and distractor may be located on different floors within the same scene, which can cause the distractor to be absent in the top-down maps.

Table 1. Comparison of GDN performance between SAVi and MAGNet on the test split under *Clean Environments*. *Sounding* and *Silent* denote metrics measured during the goal’s sound-emitting and silent periods, respectively. All metrics are macro-averaged across goal categories. Note that a low LE_{CD} accompanied by an extremely low $F_{\leq 20^\circ}$ is not informative.

Method	Sounding					Silent				
	$ER_{\leq 20^\circ} \downarrow$	$F_{\leq 20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$	$RDE \downarrow$	$ER_{\leq 20^\circ} \downarrow$	$F_{\leq 20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$	$RDE \downarrow$
SAVi	0.898	0.129	43.07	0.473	0.265	0.996	0.005	25.52	0.013	0.214
MAGNet w/o memory	0.927	0.119	24.07	0.192	0.108	0.999	0.001	33.84	0.007	0.177
MAGNet w/o self-motion	0.797	0.259	33.33	0.500	0.126	0.924	0.124	39.55	0.187	0.151
MAGNet	0.762	0.290	36.77	0.601	0.117	0.905	0.140	48.83	0.368	0.166

7. Analysis and Visualization of SELD Results

ACCDDOA format. The activity-coupled Cartesian distance and direction-of-arrival (ACCDDOA) format [9, 13] is used to represent the goal description. It is defined as:

$$\mathbf{y}_{ct} = [a_{ct}\mathbf{R}_{ct}, d_{ct}], \quad (7)$$

where c and t denote the category and time step indices, respectively. Here, $\mathbf{R}_{ct} = [x_{ct}, y_{ct}, z_{ct}] \in [-1, 1]^3$ represents the unit-norm direction-of-arrival (DOA) vector, $a_{ct} \in \{0, 1\}$ indicates the activity status of the sound event (0 for inactive and 1 for active), and $d_{ct} > 0$ is the normalized distance (scaled by 20 m). At each time step, the goal descriptions have a dimension of 84, corresponding to the 21 goal categories $\times [x_{ct}, y_{ct}, z_{ct}, d_{ct}]$.

SELD metrics. Following standard SELD practice [9, 10], we jointly evaluate detection and localization performance using: 1) the error rate and F1-score within a 20° localization tolerance ($ER_{\leq 20^\circ}$ and $F_{\leq 20^\circ}$), and 2) the localization error, localization recall, and relative distance error conditioned on correctly detected events (LE_{CD} , LR_{CD} , and RDE). Since detection and localization are evaluated jointly, the reported scores are typically lower than in separate evaluations. For example, a sound event that is correctly detected but localized outside the 20° tolerance window is counted as a false negative, whereas a misdetected event is excluded from localization evaluation even if its true location lies within the tolerance window.

Angle definitions. The azimuth ϕ_{ct} and elevation θ_{ct} angles are computed from the estimated DOA vector \mathbf{R}_{ct} as:

$$\begin{aligned} \phi_{ct} &= \arctan(y_{ct}/x_{ct}), \\ \theta_{ct} &= \arcsin(z_{ct}/\|\mathbf{R}_{ct}\|), \end{aligned} \quad (8)$$

where $\arctan(\cdot)$ and $\arcsin(\cdot)$ denote the arctangent and arcsine functions, respectively. The audio sensor is mounted at the agent’s head (with a 1.5 m height), so all angles are defined relative to the agent’s head orientation. Since the agent’s coordinate system differs from that required by the SELD metrics, a coordinate transformation is applied. In the SELD metrics convention, azimuth is zero at the front, increases counter-clockwise within $[-180^\circ, 180^\circ]$, and elevation ranges within $[-90^\circ, 90^\circ]$, increasing upward. The goal is always located 1.5 m above

the floor, which matches the sensor height. Therefore, $z_{ct} = 0$ for all cases, yielding an elevation angle $\theta_{ct} = 0^\circ$.

Sound event status. The sound event status a_{ct} is derived from the length of the estimated DOA vector \mathbf{R}_{ct} :

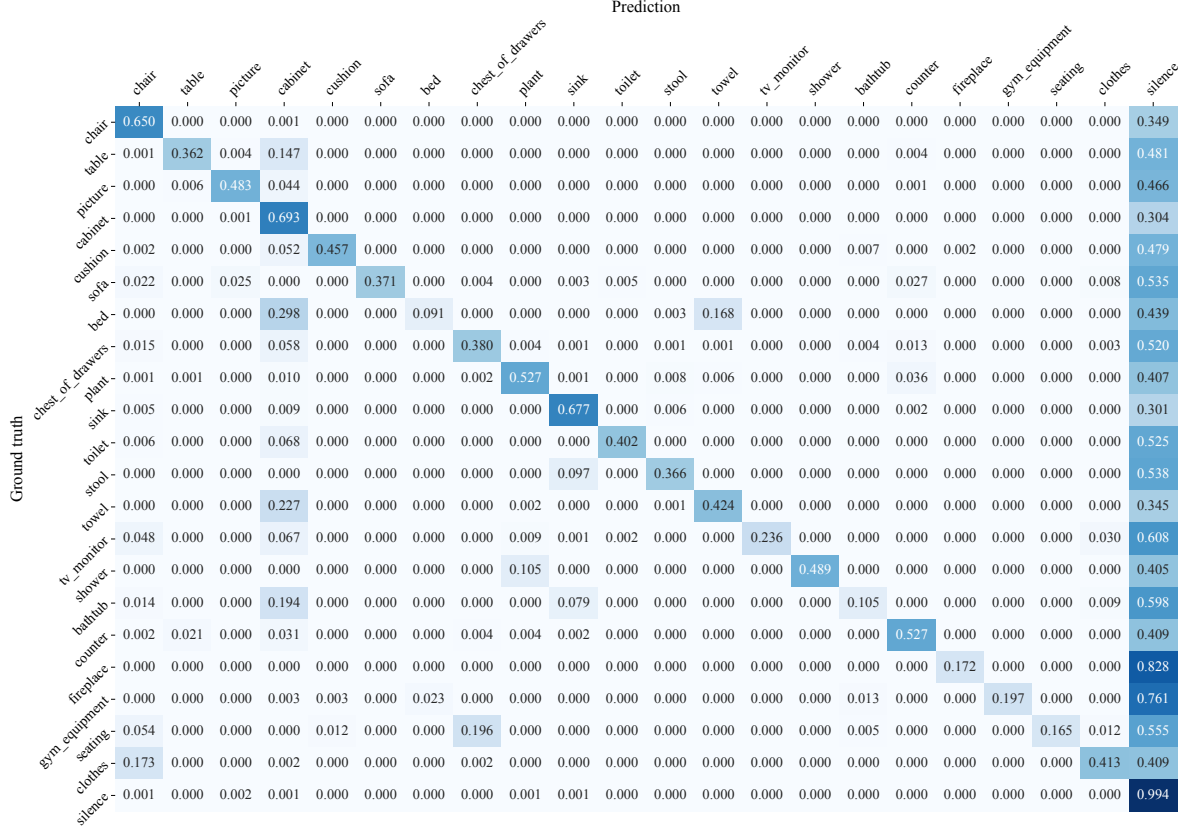
$$a_{ct} = \mathbb{I}\left(\sqrt{x_{ct}^2 + y_{ct}^2 + z_{ct}^2} \geq T_{\text{sed}}\right), \quad (9)$$

where $\mathbb{I}(\cdot)$ is the indicator function and $T_{\text{sed}} = 0.5$ serves as the activity threshold in our experiments.

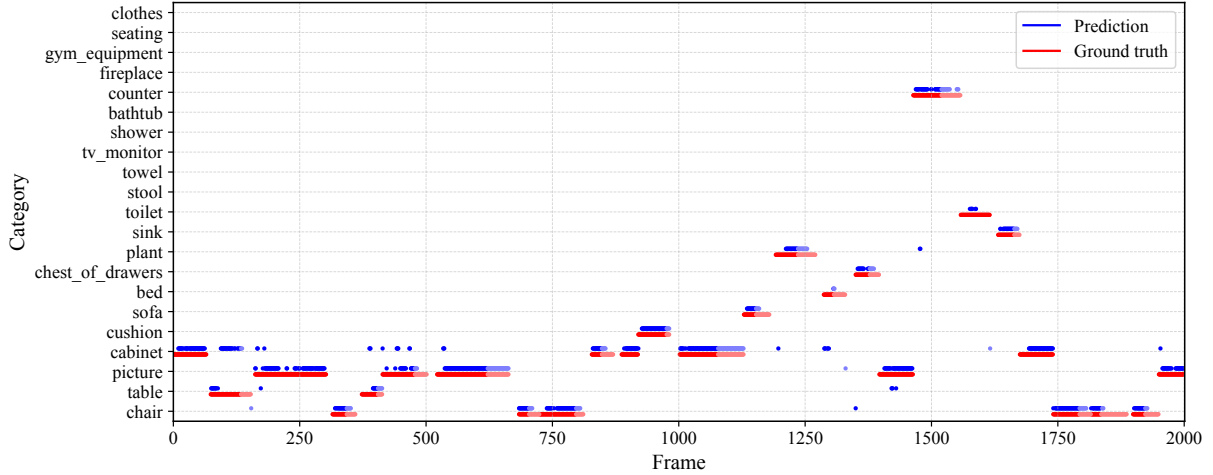
SELD results. Tab. 1 reports the SELD results of SAVi and our method with different components removed under *Clean Environments*. SAVi performs similarly to our method without the episodic memory module, as it simply aggregates goal information over time using a weighting factor λ [4]. During the goal’s sound-emitting period, our method consistently outperforms SAVi across all metrics, indicating more accurate goal detection and localization. In the silent period, the performance gap narrows but remains noticeable, consistent with the navigation results.

The second-to-last column, LR_{CD} , reports the localization recall over correctly detected events. This metric reflects how accurately a method can localize sound events that it has successfully detected. With the introduction of the episodic memory module, our method achieves a higher LR_{CD} , demonstrating that storing and retrieving historical goal information helps the model maintain more reliable spatial reasoning after the sound ceases. Furthermore, incorporating self-motion cues leads to an additional improvement in LR_{CD} , indicating that ego-motion information (e.g., pose changes and taken actions over time) provides valuable geometric constraints that enhance localization accuracy, particularly when the goal becomes silent.

Visualization. Fig. 3 and Fig. 4 show the detection and localization results of our proposed MAGNet under *Clean Environments*. Different colors indicate different goal categories, with lighter shades indicating the results during the goal’s silent period. The results are aggregated across episodes for up to 2,000 time steps (frames). The confusion matrix shown in Fig. 3a reveals that most acoustic events, regardless of category, are frequently misclassified as silence, resulting in a substantial number of false negatives. As shown in Fig. 3b, the GDN generally detects the onsets,



(a) Confusion matrix of sound event detection.

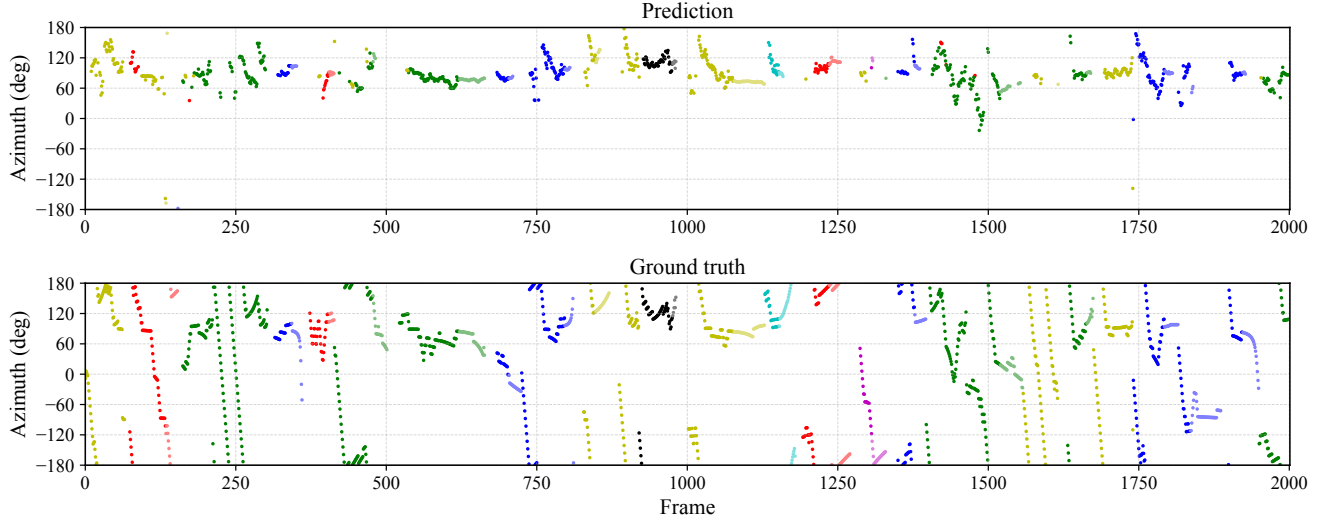


(b) Detection results of sound events a_{ct} .

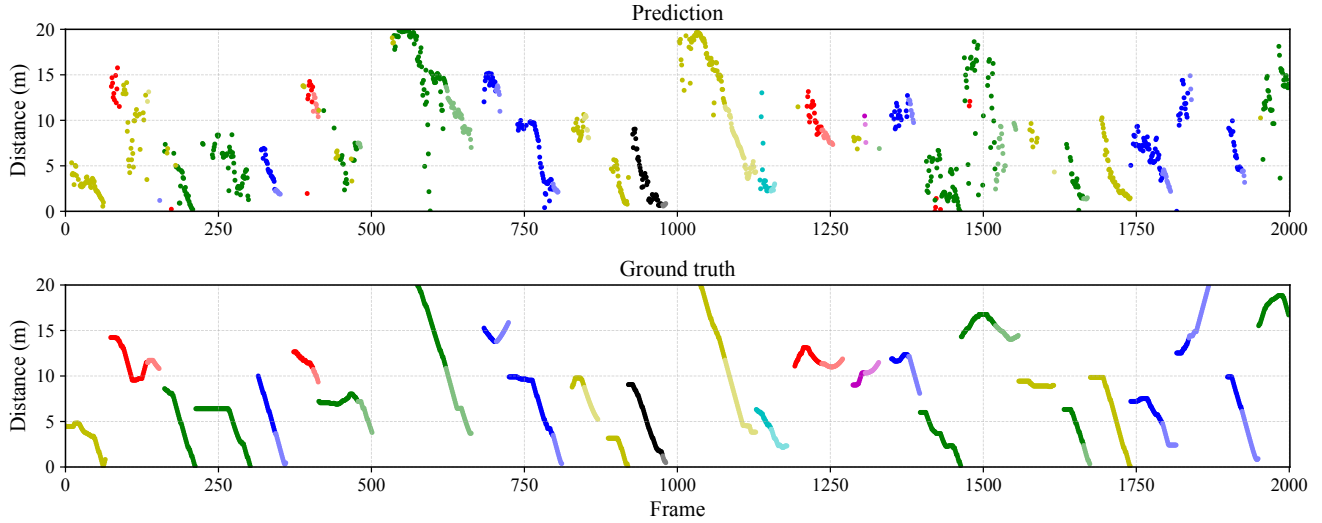
Figure 3. Visualization of the detection performance of our proposed MAGNet under *Clean Environments*.

offsets, and categories of sound events correctly. Furthermore, Fig. 4a and Fig. 4b indicate that the GDN can estimate the azimuth angle and distance of sound events, while localization errors remain noticeable. These errors are especially pronounced when the agent’s orientation changes rapidly, indicating a need for future improvements.

Limitations. Although our proposed MAGNet equipped with the memory-augmented GDN outperforms SAVi in SELD performance, its capabilities remain limited. Challenges such as complex multi-room layouts, long distances, high reverberation, unbalanced category distributions, insufficient training data, and the presence of dis-



(a) Azimuth angle ϕ_{ct} of sound event localization.



(b) Distance d_{ct} of sound event localization.

Figure 4. Visualization of the localization performance of our proposed MAGNet under *Clean Environments*.

tractor sounds continue to affect the network’s performance. Furthermore, while binaural audio captures some spatial information, its two channels make it difficult to fully represent the spatial characteristics of sound sources [16]. Most importantly, the network still struggles to generalize to unseen environments and unheard sounds, highlighting the need for further improvements to enhance its robustness and reliability. In future work, we plan to address these limitations by leveraging more diverse datasets (e.g., the BeDAViN dataset proposed in ENMuS³ [12]), employing advanced acoustic modeling techniques, and incorporating first-order Ambisonics (FOA) audio representations to better capture the spatial characteristics of sound sources.

8. Incorporating Vision for Goal Inference

Incorporating visual cues to enhance sound event localization and detection has recently attracted increasing attention. The STARSS23 [14] dataset provides multichannel audio, aligned panoramic video, and spatiotemporal annotations of sound events, facilitating research on audio-visual SELD. In this dataset, sound sources are consistently visible in the panoramic video, making audio-visual fusion effective. Some recent works have explored audio-visual fusion for SELD [1, 7].

In contrast, in the SAVN-CE task, the sound-emitting goal is often outside the agent’s field of view for most of the episode. Consequently, incorporating vision for goal in-

Table 2. Results with and without the visual encoder under *Clean Environments*.

	Navigation Metrics					SELD Metrics				
	SR \uparrow	SPL \uparrow	SNA \uparrow	DTG \downarrow	SWS \uparrow	$ER_{<20^\circ}\downarrow$	$F_{<20^\circ}\uparrow$	$LE_{CD}\downarrow$	$LR_{CD}\uparrow$	$RDE\downarrow$
MAGNet w/ visual encoder	35.1	30.8	25.4	8.5	8.6	0.827	0.227	34.45	0.501	0.131
MAGNet w/o visual encoder	37.7	32.9	27.4	8.0	10.6	0.816	0.240	31.30	0.494	0.119

Table 3. Navigation performance on SAVN-CE dataset under *Clean* and *Distracted Environments* with noise sounds at SNR = 20 dB.

	Clean Environments							Distracted Environments							
	SR↑	SPL↑	SNA↑	DTG↓	SWS↑	Reward↑	NA	SR↑	SPL↑	SNA↑	DTG↓	SWS↑	Reward↑	NA	DSR
AV-Nav	10.3	8.4	6.9	13.1	0.3	-1.07	398	9.7	8.1	5.7	14.6	0.6	-2.86	419	4.5
SMT + Audio	14.2	11.8	9.4	12.8	1.2	0.16	339	12.1	9.9	6.9	14.1	1.8	-1.19	323	4.7
SAVi	14.9	12.0	9.8	13.5	1.2	-0.47	337	14.1	11.4	8.7	12.7	2.1	0.03	360	4.2
MAGNet (Ours)	9.0	8.3	6.8	12.3	1.1	3.01	55	12.0	10.3	7.7	12.0	2.7	3.56	59	5.5

ference may provide limited benefit, and we do not use a visual encoder in MAGNet for this purpose. To validate this design choice, we compare the performance of MAGNet with and without a visual encoder (using the same structure as described in the Multimodal Observation Encoder). The experimental results in Tab. 2 indicate that adding the visual encoder while compressing the pose/audio encoder space can degrade both navigation and SELD performance. Nevertheless, more effective integration strategies may further improve overall performance.

9. Background Noise

Neither under *Clean Environments* nor *Distracted Environments* is the simulator completely silent outside the goal’s sound onsets and offsets. Therefore, we evaluate all methods under both scenarios with continuous background noise throughout each episode, without retraining. The noise sounds are sampled from the NOISEX-92 database [15] and mixed with the convolved goal sound at a fixed signal-to-noise ratio (SNR) of 20 dB.

As reported in Tab. 3, background noise noticeably degrades navigation performance, especially for models trained under *Clean Environments*. This suggests that agents exposed to distractor sounds during training have learned to better discriminate goal-related audio cues from non-goal sounds and therefore exhibit greater robustness. Among all the methods, SAVi achieves the best performance under both scenarios, demonstrating its robustness in the presence of background noise.

Although MAGNet exhibits scores lower on most metrics, it shows clear advantages in several key aspects: the smallest DTG, the highest cumulative Reward, and a notably lower NA (number of actions) compared to other methods. However, its execution of the *Stop* action is sub-optimal under noisy conditions. Unlike other methods that may continue moving when uncertain, MAGNet exhibits a tendency to stop prematurely, revealing a limitation to generalize the stopping decision appropriately under noise.

Interestingly, MAGNet performs better under *Distracted Environments* than under *Clean Environments*, which contrasts with the behavior of other methods. This can be attributed to the training conditions: in *Clean Environments*, the goal-related acoustic cues are clean and reliable, encouraging MAGNet to develop an over-reliance on these cues. Once background noise is introduced at test time, this reliance becomes problematic, as the altered acoustic conditions make the cues unreliable and lead to greater performance degradation. In contrast, training under *Distracted Environments* exposes the model to more acoustically complex conditions, preventing such over-reliance and resulting in greater robustness when noise is present.

References

- [1] Davide Berghi, Peipei Wu, Jinzheng Zhao, Wenwu Wang, and Philip JB Jackson. Fusion of audio and visual embeddings for sound event localization and detection. In *Proc. IEEE ICASSP*, pages 8816–8820, 2024. 6
- [2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *Proc. Int. Conf. 3D Vis.*, pages 667–676, 2017. 2
- [3] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicens Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-visual navigation in 3D environments. In *Proc. Eur. Conf. Comput. Vis.*, pages 17–36, 2020. 2
- [4] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15516–15525, 2021. 2, 4
- [5] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *Proc. Int. Conf. Learn. Represent.*, pages 4861–4876, 2021. 2
- [6] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. SoundSpaces 2.0:

- A simulation platform for visual-acoustic learning. In *Proc. Adv. Neural Inform. Process. Syst.*, pages 8896–8911, 2022. [1](#)
- [7] Ya Jiang, Qing Wang, Jun Du, Maocheng Hu, Pengfei Hu, Zeyan Liu, Shi Cheng, Zhaoxu Nian, Yuxuan Dong, Mingqi Cai, Xin Fang, and Chin-Hui Lee. Exploring audio-visual information fusion for sound event localization and detection in low-resource realistic scenarios. In *Proc. IEEE ICME*, pages 1–6, 2024. [6](#)
- [8] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the Nav-Graph: Vision-and-language navigation in continuous environments. In *Proc. Eur. Conf. Comput. Vis.*, pages 104–120, 2020. [2](#)
- [9] Daniel Aleksander Krause, Archontis Politis, and Annamaria Mesaros. Sound event detection and localization with distance estimation. In *Proc. EUSIPCO*, pages 286–290, 2024. [4](#)
- [10] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen. Overview and evaluation of sound event localization and detection in DCASE 2019. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 29: 684–698, 2021. [4](#)
- [11] Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Sharath Adavanne, Daniel Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, and Tuomas Virtanen. STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Proc. Detection Classification Acoust. Scenes Events Workshop*, pages 125–129, 2022. [2](#)
- [12] Zhanbo Shi, Lin Zhang, Linfei Li, and Ying Shen. Towards audio-visual navigation in noisy environments: A large-scale benchmark dataset and an architecture considering multiple sound-sources. In *Proc. AAAI*, pages 14673–14680, 2025. [2](#), [6](#)
- [13] Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, and Yuki Mitsufuji. ACCDOA: Activity-coupled Cartesian direction of arrival representation for sound event localization and detection. In *Proc. IEEE ICASSP*, pages 915–919, 2021. [2](#), [4](#)
- [14] Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A. Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Proc. Adv. Neural Inform. Process. Syst.*, pages 72931–72957, 2023. [6](#)
- [15] Andrew Varga and Herman JM Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.*, 12(3):247–251, 1993. [7](#)
- [16] Zeyuan Yang, Jiageng Liu, Peihao Chen, Anoop Cherian, Tim K Marks, Jonathan Le Roux, and Chuang Gan. RILA: Reflective and imaginative language agent for zero-shot semantic audio-visual navigation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16251–16261, 2024. [6](#)