Audio Spatially-Guided Fusion for Audio-Visual Navigation

Xinyu Zhou[1,2,3], and Yinfeng Yu[1,2,3,✉]
[1]Joint Research Laboratory for Embodied Intelligence, Xinjiang University
[2]Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Xinjiang University
[3]School of Computer Science and Technology, Xinjiang University, Urumqi 830017, China

***Abstract*—Audio-visual Navigation refers to an agent utilizing visual and auditory information in complex 3D environments to accomplish target localization and path planning, thereby achieving autonomous navigation. The core challenge of this task lies in the following: how the agent can break free from the dependence on training data and achieve autonomous navigation with good generalization performance when facing changes in environments and sound sources. To address this challenge, we propose an Audio Spatially-Guided Fusion for Audio-Visual Navigation method. First, we design an audio spatial feature encoder, which adaptively extracts target-related spatial state information through an audio intensity attention mechanism; based on this, we introduce an Audio Spatial State Guided Fusion (ASGF) to achieve dynamic alignment and adaptive fusion of multimodal features, effectively alleviating noise interference caused by perceptual uncertainty. Experimental results on the Replica and Matterport3D datasets indicate that our method is particularly effective on unheard tasks, demonstrating improved generalization under unknown sound source distributions.**

***Index Terms*—Audio-Visual Navigation, Audio-Visual Fusion, Spatial Audio, Scene Generalization**

I. Introduction

In recent years, the rapid development of emerging industries such as autonomous driving [1] and smart home systems [2] has motivated researchers to explore new navigation paradigms, among which embodied navigation [3]–[5] has attracted increasing attention. Audio-Visual Navigation (AVN) [6], [7] is a significant research direction in the field of embodied navigation. It aims to enable agents to localize targets and navigate autonomously using audio-visual observations in complex environments, with broad application potential in scenarios such as rescue and household service. Current audio-visual navigation algorithms can achieve a navigation success rate of 98% in tests with heard sounds. However, their performance drops significantly to 52% in tests with unheard sounds. This indicates that agents in current research have obvious deficiencies in generalization ability across sound sources and scenarios. Yet real application scenarios involve diverse types of sounds. We cannot achieve full coverage with training data. Therefore, improving the generalization ability of navigation models is a key issue.

Through the analysis of existing research results, we can attribute the limited generalization performance to the following two points: 1) Insufficient audio feature extraction capability, where simple convolutional encoders struggle to model temporal dependencies and fully exploit spatial cues in audio [8]–[10]. 2) Limited flexibility in multimodal fusion, where fixed-weight fusion cannot adaptively adjust the contributions of audio and visual information under changing scenes or sound categories [11], [12].

✉Yinfeng Yu is the corresponding author(E-mail: yuyinfeng@xju.edu.cn).

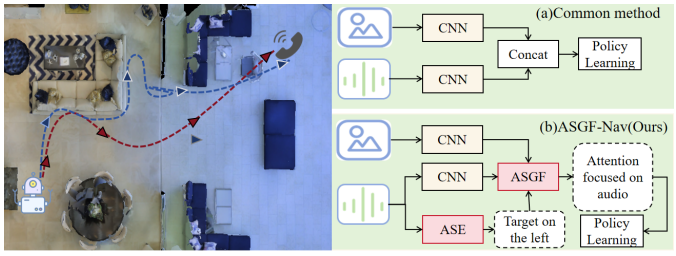

Fig. 1. Comparison of navigation trajectories and model architectures. Left: Top-down view of navigation trajectories in an indoor scene (blue: our ASGF-Nav; red: baseline method). Right: (a) Common audio-visual navigation method with simple feature concatenation; (b) Our ASGF-Nav model, which integrates the ASE module to extract spatial information from audio and uses the ASGF module to dynamically fuse audio-visual features.

In 3D environments, sound propagation is affected by factors such as relative direction, propagation distance, and environmental structure. Such space-related information is implicitly embedded in audio signals. Based on this understanding, we propose a cross-modal fusion framework guided by audio spatial states. The agent learns spatial state information from audio, and uses this information to guide the dynamic fusion of visual and audio features. In this way, the agent is no longer limited to memorizing simple correspondences between sounds and environments. Instead, it can exploit the spatial constraints provided by audio, achieving fast localization and efficient autonomous navigation of target sound sources. Experimental results show that compared with existing research, our proposed model achieves clear improvements on Unheard tasks and exhibits stronger generalization ability.

In summary, our main contributions are as follows:

- We propose a temporal-aware audio feature extraction module based on CRNN, combined with audio intensity attention, to effectively extract spatial state information from audio.
- We propose a cross-modal fusion framework guided by audio spatial states, which adaptively coordinates the weights of audio and visual modalities in different sce-

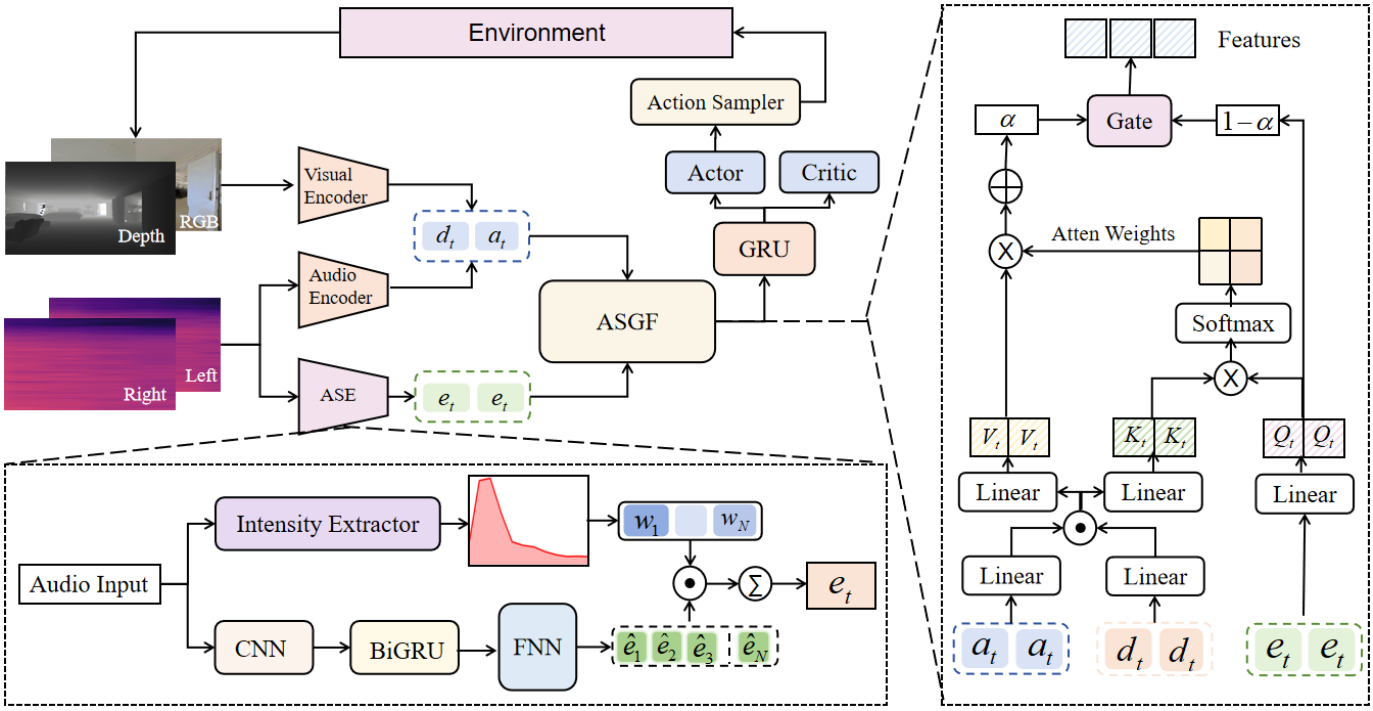

Fig. 2. Model architecture. Our audio spatially-guided fusion for audio-visual navigation model (ASGF-Nav) uses the ASE module to extract implicit spatial state information from binaural spectrograms, and guides the dynamic fusion of visual and audio features in the ASGF module to provide a basis for policy selection.

narios.

- We conduct experiments in complex 3D environments and compare with multiple baseline models, proving that the proposed method has good generalization performance.

II. Related Work

A. Audio-Visual Navigation

Audio-Visual Navigation requires agents to use visual and auditory observation information in complex 3D environments to autonomously plan paths and reach sound source targets [13]. In recent years, existing studies have explored the audio-visual navigation task from different processing methods of application scenarios and observation information. SoundSpaces [6] builds the first simulation platform for audio-visual navigation tasks based on the Replica [7] and Matterport3D [14] real indoor scene datasets, providing a unified experimental environment for audio-visual navigation methods. LLA [15] adopts a phased perception and planning structure to enhance the stability during the policy learning process. AV-WAN [16] models observation information as sound and visual maps, achieving efficient navigation by decomposing long-term goals into short-term goals. CAHM [17] combines local geometric maps with historical observations to achieve tracking of slowly moving sound sources. SAVi [18] further introduces semantic information to explicitly construct the correspondence between sound events and objects in the environment. SAAVN [19] constructs a complex acoustic environment with interference and adversarial attacks to improve the robustness of the navigation agent. ORAN [20] changes the traditional single-view observation paradigm by collecting information in an omnidirectional manner and employing knowledge distillation to achieve effective transfer of target navigation policies. Inspired by vision-language navigation, AVLEN [21] and CAVEN [22] introduce an Oracle to provide natural language instructions to the agent to assist it in locating navigation targets. More recently, AVN research has further explored collaborative acoustic perception, echo-enhanced navigation, multi-agent cooperation, and stereo-aware dynamic fusion strategies [8], [9], [23], [24]. Related embodied navigation studies have also investigated enhanced object-level perception to improve policy learning [25]. Despite these advances, how to better exploit audio information for robust navigation in complex and unseen environments remains an open problem.

B. Feature Fusion

In audio-visual navigation tasks, effective navigation depends on jointly leveraging visual and auditory modalities. From the perspective of fusion strategies, most existing approaches adopt early-stage feature-level fusion with fixed weights. Specifically, methods such as SoundSpaces [6], LLA [15], and AV-WAN [16] independently encode visual and audio features and then simply concatenate or linearly combine them as input to the policy network. However, such fusion strategies are unable to dynamically adjust the contribution of each modality in response to environmental changes or perceptual uncertainty. With the advancement of

attention mechanisms, an increasing number of works have introduced self-attention, cross-modal attention, and dynamic fusion into the fusion process [9], [11], [12], [26]. For example, FSAAVN [26] employs a self-attention module to achieve context-aware fusion of audio and visual features. SAVi [18] combines Transformer-based architectures with memory modules to address the issue of information loss under sparse or intermittent sound conditions. Related dynamic gating and end-to-end audio-visual fusion ideas have also been explored in broader audio-visual learning tasks [10], [27]. Nevertheless, agent performance still degrades significantly in the *Unheard* environment.

III. Method

We design an audio spatial state-guided cross-modal fusion framework to enhance the generalization capability of the model. As illustrated in Fig. 2, RGB and depth images are used as visual inputs, while binaural spectrograms are used as auditory inputs. Visual and auditory observations are first processed by convolutional neural networks to extract modality-specific features. For the auditory modality, we further introduce an Audio Spatial State Encoder (ASE) to explicitly learn spatial state information embedded in audio signals. The extracted visual features, audio features, and audio spatial state representations are then fed into the Audio Spatial State-Guided Cross-Modal Fusion (ASGF) module for feature integration. The fused features are recursively combined with historical hidden states through a gated recurrent unit (GRU), producing state representations enriched with temporal contextual information. The fused features are finally fed into an Actor-Critic network for policy learning and value estimation.

A. Observations Encoder

At each time step t, two independent CNNs are used as the visual and audio encoders to extract visual and auditory features, denoted by $D_t \in \mathbb{R}^{N_d}$ and $A_t \in \mathbb{R}^{N_a}$, respectively.

We introduce an *Audio Spatial State Encoder* (ASE) to model auditory observations and capture the spatial state information conveyed by audio signals, as shown in Fig. 2. Raw audio signals are first transformed into binaural spectrograms using the Short-Time Fourier Transform (STFT) [28]. The resulting spectrograms are fed into a CNN to extract hierarchical time–frequency features while preserving the temporal axis as a sequence. A BiGRU is then applied to model temporal dependencies and produce frame-wise audio spatial embeddings $\tilde{\mathbf{e}}_{t,i} \in \mathbb{R}^{N_e}$ for the current decision step t, where $i = 1, \ldots, N$ indexes the frames in the audio segment, N is the number of frames, and N_e is the embedding dimension ($N_e = 512$ in our implementation).

Meanwhile, we calculate the sound intensity of each frame in the original binaural spectrogram, normalize the sound intensity via softmax, and obtain the frame-level attention weights. Specifically, the intensity of frame i at decision step t is defined as

$$I_{t,i} = \mathrm{mean}_{f,c}\big|X_t(f,i,c)\big|, \tag{1}$$

where $X_t(f,i,c)$ denotes the magnitude at frequency bin f, frame index i, and channel c. These intensity scores are converted into temporal attention weights via a softmax along the time axis:

$$w_{t,i} = \frac{\exp\big(I_{t,i}\big)}{\sum_{j=1}^{N} \exp\big(I_{t,j}\big)}, \tag{2}$$

such that $\sum_{i=1}^{N} w_{t,i} = 1$.

Finally, the frame-wise embeddings are aggregated using the intensity-based attention weights to form a global audio spatial state representation at the current time step:

$$\mathbf{e}_t = \sum_{i=1}^{N} w_{t,i}\, \tilde{\mathbf{e}}_{t,i} \in \mathbb{R}^{N_e}, \tag{3}$$

based on the audio intensity attention mechanism, the model can focus more on frames with high confidence and obtain a more robust representation of audio spatial states.

B. Feature Fusion

We propose the Audio Spatial State Guided Cross-Modal Fusion module for adaptive multimodal fusion, as illustrated in Fig. 2.

At time step t, the audio spatial state e_t, visual feature d_t, and auditory feature a_t are first normalized and projected into a unified embedding space. The projected visual and auditory features are then concatenated to form a cross-modal memory representation, from which Key and Value representations are generated, while the audio spatial state serves as the Query. Specifically, the cross-attention computation is formulated as follows:

$$X_t = \big[W_d\,\mathrm{LN}(d_t),\ W_a\,\mathrm{LN}(a_t)\big], \tag{4}$$

$$K_t = X_t W_K, \quad V_t = X_t W_V, \tag{5}$$

$$q_t = W_q\,\mathrm{LN}(e_t), \tag{6}$$

$$c_t = \mathrm{Attn}(q_t, K_t, V_t), \tag{7}$$

where $W_d \in \mathbb{R}^{D\times N_d}$ and $W_a \in \mathbb{R}^{D\times N_a}$ project d_t and a_t into the same D-dimensional latent space; $W_q \in \mathbb{R}^{D\times N_e}$, $W_K \in \mathbb{R}^{D\times D}$, and $W_V \in \mathbb{R}^{D\times D}$ are learnable projection matrices; and c_t denotes the context representation obtained by selectively aggregating visual and auditory information conditioned on the audio spatial state.

The attended context is combined with the original audio spatial state through a residual connection:

$$\hat{h}_t = c_t + e_t, \tag{8}$$

To handle varying modality reliability, we further introduce an adaptive gating mechanism. The gating network takes e_t as input and outputs a scalar coefficient through a multi-layer perceptron (MLP):

$$\alpha_t = \sigma\big(\mathrm{MLP}(\mathrm{LN}(e_t))\big), \quad \alpha_t \in (0,1), \tag{9}$$

where $\sigma(\cdot)$ denotes the Sigmoid function. The final fused state is obtained as

$$h_t = \mathrm{LN}_{\mathrm{out}}\big(\alpha_t \hat{h}_t + (1-\alpha_t) e_t\big). \tag{10}$$

TABLE I
NAVIGATION PERFORMANCE COMPARISON WITH OTHER METHODS.

Method	Replica						Matterport3D					
	Heard			Unheard			Heard			Unheard		
	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑
Random Agent [16]	4.9	18.5	1.8	4.9	18.5	1.8	2.1	9.1	0.8	2.1	9.1	0.8
Direction Follower [16]	54.7	72.0	41.1	11.1	17.2	8.4	32.3	41.2	23.8	13.9	18.0	10.7
Gan et al. [15]	57.6	83.1	47.9	7.5	15.7	5.7	22.8	37.9	17.1	5.0	10.2	3.6
AV-WAN [16]	86.6	98.7	70.7	34.7	52.8	27.1	72.3	93.6	54.8	40.9	56.7	30.6
AGSA [9]	80.1	95.6	52.0	36.6	48.3	22.4	57.8	72.1	34.3	26.2	36.5	13.1
SoundSpaces [6]	74.4	91.4	48.1	34.7	50.9	16.7	54.3	67.7	31.3	25.9	40.5	12.8
ASGF-Nav (Ours)	82.7	94.5	62.7	**63.3**	**76.5**	**36.9**	59.1	87.6	39.5	**52.2**	**66.4**	29.9

IV. EXPERIMENTS

A. Experimental Setup

We conduct experiments on the public datasets Replica [7], Matterport3D [14], and the SoundSpaces audio simulation dataset. The Replica dataset [7] contains 18 accurately scanned 3D indoor scenes, including apartments and offices. The Matterport3D [14] dataset consists of 85 real-world indoor scenes, including residential environments. The audio perceived by the agent is synthesized by applying binaural room impulse responses, corresponding to the sound source directions, to the original sound signals. In each episode, a scene is randomly selected, along with random starting positions for the agent and the target sound source. During navigation, the agent can execute four discrete actions: {*turn right*, *turn left*, *stop*, *move forward*}.

We evaluate the model under two task settings:

1) Heard-Sound, where the target sound source remains the same (telephone) during training, validation, and testing;
2) Unheard-Sound, where different sounds are used during training and testing to evaluate the model's generalization ability to unseen sounds. In both task settings, all test scenes are unseen during training.

B. Evaluation Metrics

In the experiments, we evaluate the navigation performance of different models using three metrics:

1) Success weighted by Number of Actions (SNA) [29]: the number of actions in successful navigation episodes is weighted for evaluating action efficiency during navigation;
2) Success Rate (SR): the proportion of episodes in which the agent successfully navigates to the target during testing;
3) Success weighted by Path Length (SPL): which weights successful navigation episodes based on the deviation between the actual traveled path and the shortest path.

TABLE II
ABLATION STUDY OF ASGF-NAV COMPONENTS ON THE *Unheard* TASK.

Method	Replica (Unheard)			Matterport3D (Unheard)		
	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑
SoundSpaces	34.7	50.9	16.7	25.9	40.5	12.8
w/o ASGF	38.8	53.6	21.4	31.6	40.0	20.7
w/o ASE	59.5	71.9	34.8	41.4	59.9	22.7
ASGF-Nav	**63.3**	**76.5**	**36.9**	**52.2**	**66.4**	**29.9**

C. Comparison with Baselines

The results of the comparative experiments are shown in Table I. In the Heard task on the Replica [7] and Matterport3D [14] datasets, our model outperforms the SoundSpaces [6] baseline on all metrics, indicating that the model has stable and effective navigation ability in the Heard task. In the more challenging Unheard task, our model shows significant advantages. Especially in the Replica [7] scenarios, compared with the optimal method, the SPL, SR and SNA are improved by 28.6%, 23.7% and 9.8% respectively. In the more complex Matterport3D [14] dataset, the SR and SPL also achieve obvious improvements. The above results show that the proposed model can still achieve efficient navigation under unknown sound conditions. Its performance improvement does not come from the model's memory of specific sound directions and scenes, but from mining the spatial state information existing in the audio and realizing robust cross-modal fusion on this basis, thus reflecting good generalization ability under different environments and sound distributions.

D. Ablation Study

In our study, we conducted ablation experiments on the Unheard task of both datasets, with the model evaluated under the following three settings:

1) Removing both the ASE module and the ASGF module;
2) Removing the ASE module and using a learnable vector as the query for cross-modal fusion;

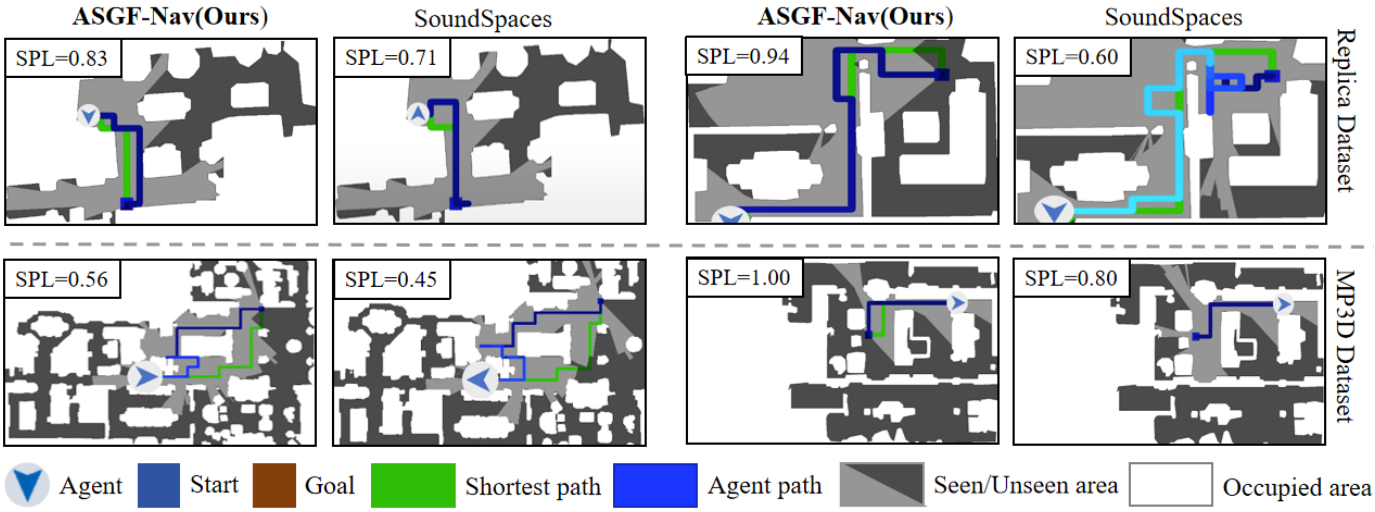

Fig. 3. Top-down visualization of agent trajectories under the Unheard task. The color gradient from dark to light blue represents temporal progression. Compared with SoundSpaces, our method reaches the target with fewer detours.

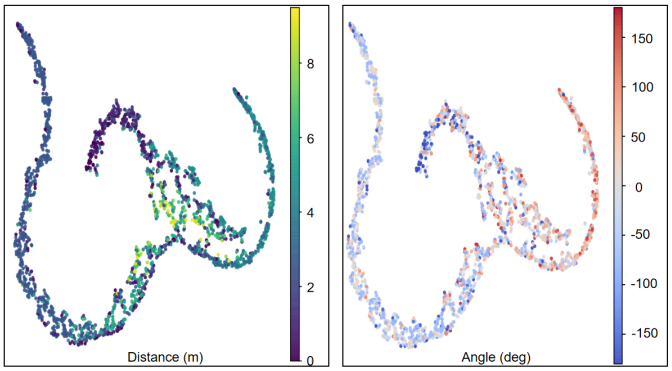

Fig. 4. t-SNE projection of the audio features extracted by the ASE module. The left figure is colored according to the distance to the target, while the right figure is colored according to the relative angle to the target.

3) Removing the ASGF module, and using the spatial state information learned by ASE together with the visual and audio features extracted by the convolutional neural network as the decision input.

The quantitative results are summarized in Table II.

It can be observed that after removing the ASGF module, the SNA, SR and SPL all decrease significantly, which indicates that cross-modal spatial guidance and feature fusion play an important role in navigation decision-making under unknown sound conditions. Under the Unheard task, removing the ASE module makes it difficult for the agent to adapt to unknown sound source distributions, leading to a significant degradation in performance, which demonstrates that the effective utilization of spatial state information in audio is an important prerequisite for stable navigation in the Unheard scenario. Overall, the ablation experiments show that the ASE module extracts spatial state information from audio to provide key geometric constraints for navigation, while the ASGF module effectively fuses this information with visual observations, thus enabling the model to maintain stable and efficient navigation performance under unknown sound conditions. This verifies the good generalization ability of the proposed method under both unseen environments and unseen sound settings.

E. Visualization Analysis

To further analyze the representations learned by the ASE module, we performed t-SNE visualization on its output high-dimensional representations, as shown in Fig. 4. The figure displays 2000 samples extracted from the Unheard task on the Replica dataset [7]. The left subfigure is colored according to the distance to the target. The right subfigure is colored according to the relative azimuth to the target. We can observe that the outputs of the ASE module do not show a random distribution in the embedding space. Instead, they form distinct continuous structures. When colored by the target distance, the samples exhibit a smooth color gradient along these structures. This suggests that the extracted features can continuously reflect the target distance information. When colored by the relative azimuth, the same continuous structures show consistent angular variation patterns. This suggests that the direction information is also effectively encoded. The above phenomena show that the ASE module learns target-related latent audio representations that preserve spatially informative cues, which can provide useful guidance for cross-modal fusion and navigation decision-making under unknown sound conditions.

We compared the top-down trajectories of ASGF-Nav and SoundSpaces in the 3D environment, as shown in Fig. 3. From the trajectory plots, we can see that the navigation paths of ASGF-Nav are generally closer to the shortest path compared with the baseline. This result corresponds to its advantages in the SPL metric. In the more complex Matterport3D dataset [14], this difference becomes even more obvious. The trajectories of SoundSpaces [6] often deviate

from the optimal route. They also involve multiple direction corrections. In contrast, ASGF-Nav is more direct. It reduces unnecessary exploration processes. The trajectory comparison indicates that our method can achieve reasonable path planning even in complex environments. It also demonstrates better generalization ability.

V. Conclusion

We propose ASGF-Nav, a novel audio-visual navigation framework that leverages audio-derived spatial cues to guide dynamic multimodal fusion for improved generalization. Specifically, an Audio Spatial State Encoder is introduced to focus on reliable audio frames and learn informative latent audio representations, while an Audio Spatial State Guided Cross-Modal Fusion module adaptively balances auditory and visual contributions during decision-making. In addition, a gating mechanism is employed to suppress unreliable perceptual signals and enhance decision stability. Experimental results show that ASGF-Nav achieves superior navigation accuracy and path efficiency in complex unseen environments, demonstrating stronger adaptability and generalization. Overall, the results suggest that incorporating task-relevant audio representations into a reliable multimodal fusion mechanism is effective for audio-visual navigation.

ACKNOWLEDGMENT

This research was financially supported by the National Natural Science Foundation of China (Grant No.: 62463029).

References

[1] Y. Lin, Z. Lin, H. Chen, P. Pan, C. Li, S. Chen, K. Wen, Y. Jin, W. Li, and X. Ding, "Jarvisir: Elevating autonomous driving perception with intelligent image restoration," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 369–22 380.

[2] S. M. H. Anik, X. Gao, H. Zhong, X. Wang, and N. Meng, "Programming of automation configuration in smart home systems: Challenges and opportunities," *ACM Transactions on Software Engineering and Methodology*, 2025.

[3] H. Wang, W. Liang, L. V. Gool, and W. Wang, "Towards versatile embodied navigation," *Advances in neural information processing systems*, vol. 35, pp. 36 858–36 874, 2022.

[4] Y. Liu, L. Liu, Y. Zheng, Y. Liu, F. Dang, N. Li, and K. Ma, "Embodied navigation," *Science China Information Sciences*, vol. 68, no. 4, pp. 1–39, 2025.

[5] D. Zheng, S. Huang, L. Zhao, Y. Zhong, and L. Wang, "Towards learning a generalist model for embodied navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 624–13 634.

[6] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *European conference on computer vision*, 2020, pp. 17–36.

[7] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[8] Y. Yu, L. Cao, F. Sun, C. Yang, H. Lai, and W. Huang, "Echo-enhanced embodied visual navigation," *Neural Computation*, vol. 35, no. 5, pp. 958–976, 2023.

[9] J. Li, Y. Yu, L. Wang, F. Sun, and W. Zheng, "Audio-guided dynamic modality fusion with stereo-aware attention for audio-visual navigation," in *International Conference on Neural Information Processing*, 2025, pp. 346–359.

[10] Y. Yu, Z. Jia, F. Shi, M. Zhu, W. Wang, and X. Li, "Weavenet: End-to-end audiovisual sentiment analysis," in *International Conference on Cognitive Systems and Signal Processing*, 2021, pp. 3–16.

[11] Y. Yu, H. Zhang, and M. Zhu, "Dynamic multi-target fusion for efficient audio-visual navigation," *arXiv preprint arXiv:2509.21377*, 2025.

[12] H. Zhang, Y. Yu, L. Wang, F. Sun, and W. Zheng, "Iterative residual cross-attention mechanism: An integrated approach for audio-visual navigation tasks," *arXiv preprint arXiv:2509.25652*, 2025.

[13] Z. Shi, L. Zhang, L. Li, and Y. Shen, "Towards audio-visual navigation in noisy environments: A large-scale benchmark dataset and an architecture considering multiple sound-sources," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 14, 2025, pp. 14 673–14 680.

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

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

[16] C. Chen, S. Majumder, Z. Al-Halah, R. Gao, S. K. Ramakrishnan, and K. Grauman, "Learning to set waypoints for audio-visual navigation," in *International Conference on Learning Representations (ICLR)*, 2021.

[17] A. Younes, D. Honerkamp, T. Welschehold, and A. Valada, "Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 928–935, 2023.

[18] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 516–15 525.

[19] Y. Yu, W. Huang, F. Sun, C. Chen, Y. Wang, and X. Liu, "Sound adversarial audio-visual navigation," in *International Conference on Learning Representations*, 2022.

[20] J. Chen, W. Wang, S. Liu, H. Li, and Y. Yang, "Omnidirectional information gathering for knowledge transfer-based audio-visual navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 993–11 003.

[21] S. Paul, A. Roy-Chowdhury, and A. Cherian, "Avlen: Audio-visual-language embodied navigation in 3d environments," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6236–6249, 2022.

[22] X. Liu, S. Paul, M. Chatterjee, and A. Cherian, "Caven: An embodied conversational agent for efficient audio-visual navigation in noisy environments," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 4, 2024, pp. 3765–3773.

[23] Y. Yu, C. Chen, L. Cao, F. Yang, and F. Sun, "Measuring acoustics with collaborative multiple agents," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 335–343.

[24] H. Zhang, Y. Yu, L. Wang, F. Sun, and W. Zheng, "Advancing audio-visual navigation through multi-agent collaboration in 3d environments," in *International Conference on Neural Information Processing*, 2025, pp. 502–516.

[25] Y. Yu and D. Yang, "Dope: Dual object perception-enhancement network for vision-and-language navigation," in *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 2025, pp. 1739–1748.

[26] Y. Yu, L. Cao, F. Sun, X. Liu, and L. Wang, "Pay self-attention to audio-visual navigation," in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, 2022, p. 46.

[27] Y. Yu and S. Sun, "Dgfnet: End-to-end audio-visual source separation based on dynamic gating fusion," in *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 2025, pp. 1730–1738.

[28] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[29] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva *et al.*, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.