Spatial-Aware Conditioned Fusion for Audio-Visual Navigation

Shaohang Wu[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 tasks require agents to locate and navigate toward continuously vocalizing targets using only visual observations and acoustic cues. However, existing methods mainly rely on simple feature concatenation or late fusion, and lack an explicit discrete representation of the target's relative position, which limits learning efficiency and generalization. We propose Spatial-Aware Conditioned Fusion (SACF). SACF first discretizes the target's relative direction and distance from audio-visual cues, predicts their distributions, and encodes them as a compact descriptor for policy conditioning and state modeling. Then, SACF uses audio embeddings and spatial descriptors to generate channel-wise scaling and bias to modulate visual features via conditional linear transformation, producing target-oriented fused representations. SACF improves navigation efficiency with lower computational overhead and generalizes well to unheard target sounds.**

***Index Terms*—Spatial perception, conditional fusion, localization descriptor, Audio-visual navigation**

I. Introduction

Audio-visual navigation (AVN) utilizes audio-visual cues to localize sound sources [1], [2]. Existing methods predominantly employ simple feature concatenation [1], [3] or late fusion, as illustrated in Fig. 1. However, such fusion approaches often overlook two core issues: First, they lack explicit spatial representation. Many methods treat audio more as a directional cue or implicit hint, requiring the policy network to learn the target's location autonomously within a high-dimensional continuous feature space [4]. Second, visual representation learning lacks auditory-conditioned guidance. Visual encoders typically treat the entire field of view uniformly, struggling to dynamically emphasize target-relevant visual information based on sound source cues [5], [6]. These shortcomings limit model learning efficiency and result in poor generalization when testing in unfamiliar environments or without sound.

Human navigation in the real world first relies on binaural auditory differences to roughly determine the direction and distance of sound sources. This spatial intent then guides visual attention to search for paths in specific directions. Inspired by this, this paper proposes a novel framework called Spatial-Aware Conditioned Fusion (SACF). It aims to address the aforementioned challenges through explicit spatial discretization representation and deep cross-modal modulation. SACF comprises two core components: 1) Spatially Discretized Localization Descriptor (SDLD). Unlike traditional regression-based predictions, this module discretizes the relative position of sound sources into intervals of direction and distance, predicts their probability distributions, and encodes them into compact geometric descriptors. This approach not only reduces learning complexity but also provides robust spatial priors for subsequent decision-making. 2) To achieve deep auditory guidance for visual processing, we propose Audio-Descriptor Conditioned Visual Fusion (ACVF). This mechanism avoids simple feature concatenation by dynamically generating per-channel scaling and bias parameters, conditioned on audio embeddings and the generated SDLD. These parameters transform visual feature maps through linear modulation, thereby suppressing irrelevant background noise and enhancing target-oriented visual representations during feature extraction.

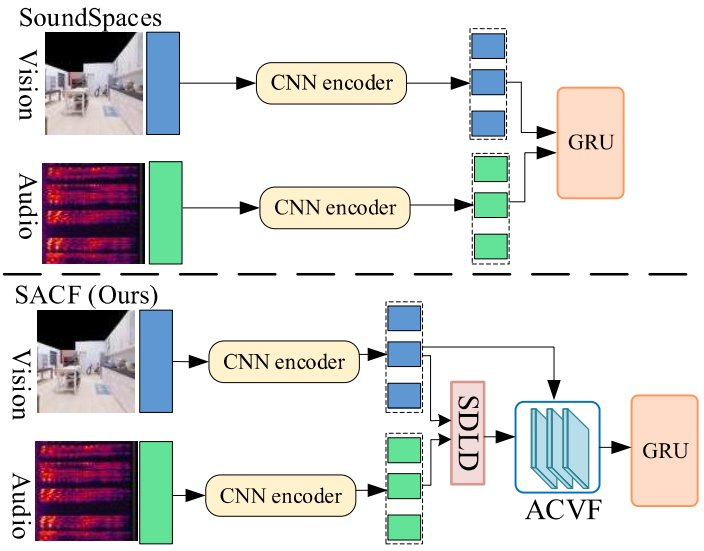

Fig. 1. Comparison between the SoundSpaces baseline and our method.

Experimental results demonstrate that SACF not only enhances overall navigation performance but also significantly improves learning efficiency and cross-scenario generalization capabilities. Specifically, ACVF employs FiLM-style channel modulation [7] instead of spatial weighting, achieving efficient cross-modal fusion with lower computational and parameter costs. Ablation studies further validate the contributions of SDLD spatial priors and ACVF modulation mechanisms to

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

performance gains.

The main contributions of this paper are summarized as follows:

- We designed the SDLD module to model the relative positions of sound sources in a direction-distance distribution format and encode them into compact descriptors for policy conditioning and state modeling.
- We proposed the ACVF mechanism, which dynamically modulates visual features based on spatial intent, significantly enhancing the directionality and effectiveness of visual representations in navigation tasks.
- We evaluated this approach on SoundSpaces, achieving favourable navigation efficiency on both the Replica and Matterport3D datasets, it exhibited exceptional generalization capabilities particularly in cross-scene tests where the target sound is unheard.

II. Related Work

A. Audio-Visual Navigation Integration Mechanism

Audio-visual navigation (AVN) is gaining increasing attention as a key task in embodied intelligence [8]. SoundSpaces [1] provides a standardized simulation platform and benchmark. Early approaches typically concatenate audio and visual features and feed them into a recursive policy [1], but this leads to limited cross-modal interactions. Subsequent work has advanced AVN primarily through two directions: fusion and memory modeling. Regarding fusion, attention-based approaches like FSAAVN [9] dynamically adjust modal weights [10]. For memory, structured acoustic memory or explicit history aggregation helps preserve sound source cues over longer time scales [11]. Additionally, some approaches incorporate active acoustic emission and echo analysis to handle sparse or missing sources [12], [13], while others leverage multi-agent collaborative acoustic measurements to enhance estimation accuracy [14]. However, existing methods still lack explicit spatial structure representation and auditory-guided visual learning. To address these limitations, our SACF significantly enhances learning efficiency and generalization capabilities through explicit spatial discretization and channel-level conditional modulation.

B. Channel-Wise Conditional Modulation

FiLM (Feature-Level Linear Modulation) was initially developed for conditional representation learning in visual reasoning tasks. This model generates channel-level scaling and bias parameters (γ, β) from conditional inputs, then modulates intermediate features via $\gamma \odot x + \beta$ to achieve lightweight conditional feature selection [7]. We observe that this channel-level modulation effectively adjusts feature weights with low overhead and good trainability. In contrast, attention-based cross-modal fusion typically relies on spatial weights or token-level interactions [9], [15]–[17], which increase computational/parameter overhead and may complicate optimization in reinforcement learning. FiLM-style modulation avoids explicit spatial reweighting, instead performing conditional selection at the channel dimension. This makes it more concise and efficient, better suited for fusion tasks in reinforcement learning [7], [18].

III. Methodology

As shown in Fig. 2, this paper proposes the Spatial-Aware Conditioned Fusion (SACF) framework for audio-visual navigation tasks. The overall process consists of three stages: perception, decision, and action. During the perception stage, the agent interacts with the environment to obtain first-person multimodal observations $o_t = \{I_t, D_t, A_t\}$ at each time step t, where I_t and D_t represent RGB and depth images, respectively, and A_t denotes the audio observation. During the decision phase, the visual and auditory information obtained from observations are processed through visual and auditory encoders, respectively, yielding visual feature maps F_t^v and audio feature maps F_t^a. Subsequently, Spatially Discretized Localization Descriptor (SDLD) module processes these to yield the spatial descriptor vector g_t. Building upon this, the Audio-Descriptor Conditioned Visual Fusion (ACVF) module generates FiLM-style channel modulation parameters conditioned on $[F_t^a; g_t]$, producing a more directional fused visual representation. Finally, the fused representation feeds into a Gated Recurrent Unit (GRU) [19] to model temporal state, and through Actor-Critic [20], [21] outputs action (a_t) and value estimates. The agent continuously interacts with the environment based on these outputs until the vocalization goal is achieved. The following sections detail the structural design and implementation specifics of SDLD and ACVF respectively.

A. Spatially Discretized Localization Descriptor

Spatially Discretized Localization Descriptor (SDLD) explicitly infers the relative spatial intent of sound sources from joint visual and auditory cues. It discretizes the target's relative spatial position into direction and distance, predicts its distribution, and encodes it into a compact descriptor for use in policy conditioning and state modeling. First, to obtain a joint representation for localization, we fuse the visual feature map F_t^v with the audio feature map F_t^a to produce the audio-visual joint feature F_{av}. As shown in Fig. 3.

Direct regression on continuous coordinates often struggles to capture the multimodal spatial distribution caused by echoes in real environments and can easily lead to training oscillations. In contrast, discretized representations can explicitly model localization uncertainty through probability distributions. Based on this, we divide the continuous space into $K = 20$ discrete categories, a choice driven by computational efficiency and physical granularity. For a 30-meter range, 20 intervals correspond to an average interval size of 1.5 meters. This is roughly equivalent to the typical step size of an agent or the safe obstacle-avoidance distance in indoor navigation, providing sufficient spatial resolution without introducing excessive gaps in the distribution. Distance classification covers the range $0 \sim 30$ meters, while angle classification spans the full range $-\pi \sim +\pi$. The position prediction head outputs

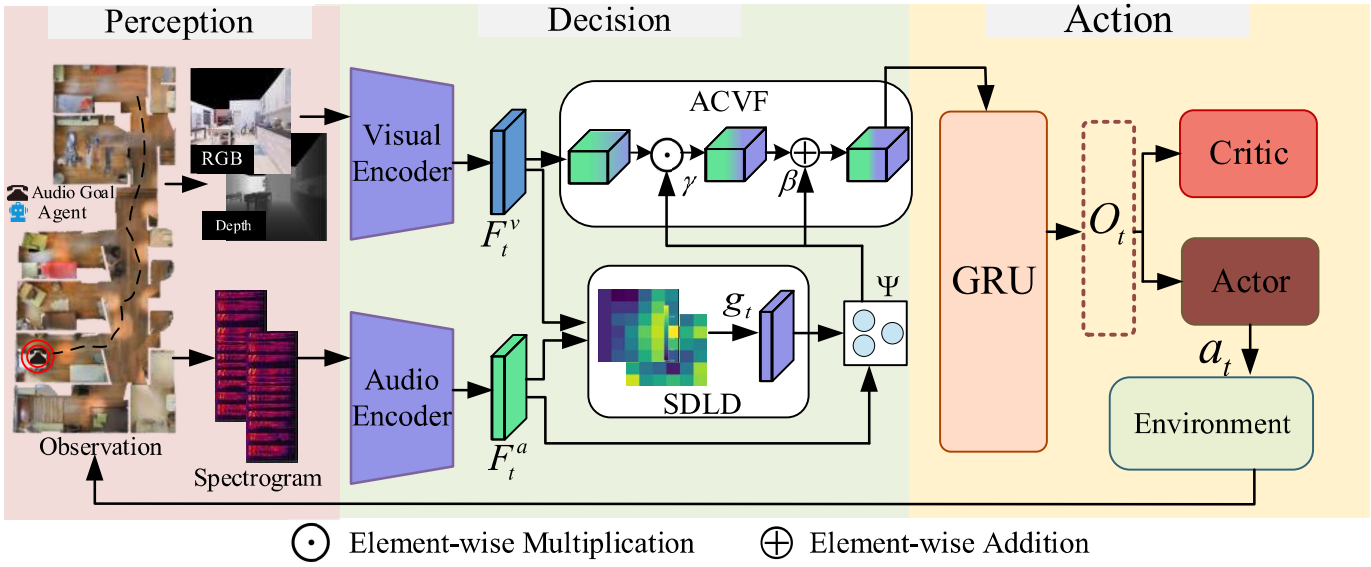

Fig. 2. SACF for audio-visual navigation receives image and audio observations $o_t = \{I_t, D_t, A_t\}$ at each time step. Features are extracted via visual/audio encoders, and spatial descriptors g_t are generated by SDLD. Subsequently, ACVF conditions on $[F_t^a; g_t]$ to generate FiLM-style channel modulation parameters that enhance the directionality of visual representations. These fused features are input into a GRU, which outputs action a_t and value estimates via Actor-Critic for closed-loop decision-making.

non-normalized probabilities for each category, which are then transformed via the Softmax function into smooth probability distributions P_d and P_θ. To achieve more precise continuous value estimates than simple classification, we do not directly select the category center with the highest probability. Instead, we compute the mathematical expectation of the prediction distribution using a weighted average. Let c_i^d and c_i^θ denote the center values for the ith distance and angle category, respectively. The predicted distance $\hat{d}$ and azimuth $\hat{\theta}$ are then:

$$\hat{d} = \sum_{i=1}^{K} P_d(i) \cdot c_i^d, \quad \hat{\theta} = \sum_{i=1}^{K} P_\theta(i) \cdot c_i^\theta. \tag{1}$$

Subsequently, we transform the polar coordinate predictions into direction vectors in Cartesian coordinates to avoid numerical instability caused by angular periodicity:

$$x_t = \cos(\hat{\theta}), \quad y_t = \sin(\hat{\theta}). \tag{2}$$

In addition to position coordinates, the module computes a Sound Event Detection (SED) score s_t to determine if the agent has reached the target (e.g., when $\hat{d}$ falls below a threshold). We feed the triplet (s_t, x_t, y_t) as the current observation input into a Long Short-Term Memory (LSTM) network [22], [23]. The LSTM retains memory of historical observation sequences, dynamically refining the current localization guess using temporal cues (such as the gradient of sound intensity change after the agent moves one step). It ultimately outputs a spatial descriptor g_t that incorporates spatio-temporal consistency.

B. Audio-Descriptor Conditioned Visual Fusion

To achieve deep auditory guidance of visual perception, we propose the ACVF module. Rather than simply concatenating

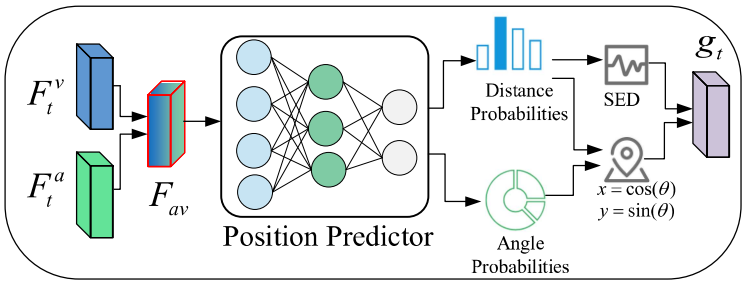

Fig. 3. SDLD fuses F_t^v and F_t^a into F_{av} to predict distance/angle distributions, then takes the expectation to obtain $\hat{d}, \hat{\theta}$. These are encoded as compact spatial descriptors using $(\cos\hat{\theta}, \sin\hat{\theta})$.

features, this module dynamically filters visual information using the Feature Linear Modulation (FiLM) [9] mechanism. It conditions on the spatial descriptor g_t generated by SDLD and the global audio feature F_t^a. Unlike spatial attention mechanisms focused on pixel-level alignment, we employ FiLM [9] for channel-level feature modulation. This mechanism dynamically adjusts the weights of visual channels through linear transformations, prioritizing the enhancement of sensitivity to key feature types such as geometric structures and passability based on auditory cues, rather than being confined to specific pixel locations. This macro-level guidance aligns closely with the navigation task's decision logic of first determining direction followed by fine-grained approach. It enables agents to rapidly identify viable path types during initial stages. Simultaneously, the minimal parameter count of channel modulation significantly reduces computational overhead, providing an efficient lightweight solution for long-horizon audio-visual reinforcement learning training. The conditional

TABLE I
PERFORMANCE COMPARISON WITH OTHER METHODS UNDER THE DEPTH SETTING. SPL, SR, SNA ARE PERCENTAGES.

Method	Replica						Matterport3D					
	Heard			Unheard			Heard			Unheard		
	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑
Random [11]	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 [11]	54.7	72.0	41.1	11.1	17.2	8.4	32.3	41.2	23.8	13.9	18.0	10.7
Frontier Waypoints [11]	44.0	63.9	35.2	6.5	14.8	5.1	30.6	42.8	22.2	10.9	16.4	8.1
Supervised Waypoints [11]	59.1	88.1	48.5	14.1	43.1	10.1	21.0	36.2	16.2	4.1	8.8	2.9
Gan et al. [2]	57.6	83.1	47.9	7.5	15.7	5.7	22.8	37.9	17.1	5.0	10.2	3.6
SoundSpaces [1]	74.4	91.4	48.1	34.7	50.9	16.7	54.3	67.7	31.3	21.9	33.5	10.4
AGSA [6]	75.5	93.2	**52.0**	36.6	48.3	**22.4**	54.1	70.0	30.0	26.2	36.5	13.1
SACF (Ours)	**80.3**	**96.3**	51.7	**43.9**	**79.1**	18.7	**55.0**	**69.0**	**32.5**	**42.4**	**58.3**	**28.0**

TABLE II
PERFORMANCE COMPARISON WITH OTHER METHODS UNDER THE RGB SETTING. SPL, SR, SNA ARE PERCENTAGES.

Method	Replica						Matterport3D					
	Heard			Unheard			Heard			Unheard		
	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑
SoundSpaces [1]	62.6	72.1	31.5	24.9	35.6	14.1	44.7	64.3	22.0	20.4	30.4	7.7
SACF (Ours)	**73.5**	**90.1**	**38.3**	**36.0**	**62.2**	**14.2**	**45.6**	**67.7**	**23.2**	**39.6**	**57.7**	**22.8**

vector $c_t = [F_t^a; g_t]$ is input into the parameter generation network Ψ to generate the corresponding scaling coefficients γ and bias coefficients β for each individual channel of the visual feature map F_t^v. Here, Ψ is simply implemented as a standard multi-layer perceptron (MLP) to effectively learn these modulating parameters:

$$\gamma, \beta = \Psi(c_t). \tag{3}$$

After spatial dimension broadcasting, these parameters perform a per-channel affine transformation on the visual features:

$$\tilde{F}_t^v = (1 + \gamma) \odot F_t^v + \beta. \tag{4}$$

ACVF is a channel-based conditional mechanism rather than a spatial attention mechanism. γ and β are broadcast across spatial positions. This mechanism amplifies channels supporting the intention while suppressing less relevant channels, thereby generating a more directional fusion representation $\tilde{F}_t^v$ for the policy. This assists the agent in making more efficient decisions.

IV. EXPERIMENTS

A. Experimental Setup

We utilize Soundspaces [1] to evaluate our method, which comprises two independent environments: Matterport3D [24] and Replica [25]. To align with prior work [1], we partition scenes into three groups: training/validation/test, consisting of 9/4/5 for Replica and 73/11/18 for Matterport3D. We evaluated our agent's performance across both environments under two distinct settings [1]: Heard and Unheard. The Unheard setting in Matterport3D is the most demanding and our primary focus. The strategy is trained from scratch. Our agent is optimized using Proximal Policy Optimization (PPO) across 5 parallel environments for 40,000 updates. We set the PPO hyperparameters as follows: clip parameter 0.1, 4 epochs per update, 1 mini-batch, value loss coefficient 0.5, entropy coefficient 0.20, and maximum gradient norm 0.5. The learning rate is 2.5×10^{-4} with linear decay, and we use $\epsilon = 1 \times 10^{-5}$.

TABLE III
ABLATION STUDIES OF THE SDLD AND ACVF MODULES ON THE UNHEARD SETTING.

Model	Replica		Matterport3D	
	SR (↑)	SPL (↑)	SR (↑)	SPL (↑)
w/o SDLD and w/o ACVF	50.9	34.7	33.5	21.9
w/o SDLD	66.1	38.6	56.2	37.8
w/o ACVF	67.6	36.7	56.2	39.5
SACF (Ours)	**79.1**	**43.9**	**58.3**	**42.4**

Note: "w/o" denotes without.

We adopt standard SoundSpaces metrics [1] for evaluation: Success Rate (SR) measures the proportion of successful navigation; Success weighted by Path Length (SPL) evaluates path efficiency; and Success weighted by Normalized Align-

ment (SNA) assesses the directional alignment between agent actions and the sound source.

TABLE IV
PARAMETER COMPARISON OF DIFFERENT AUDIO-VISUAL FUSION MECHANISMS.

Model	Replica	Matterport3D
	Params (M)↓	Params (M)↓
Simple Concatenation	**4.35M**	4.95M
Cross-Modal Spatial Attention	7.06M	7.67M
SACF (Ours)	4.50M	**4.49M**

B. Comparison Results

We compare our approach with baselines for audio-visual navigation. Table I shows that our SACF framework, using depth maps as the visual input, achieves higher SPL and SR on the Replica dataset [25] in both the Heard 80.3%/96.3% and Unheard 43.9%/79.1% splits. Compared to SoundSpaces, SACF improves SPL/SR by +5.9%/+4.9% on Heard and +9.2%/+28.2% on Unheard. On the SNA metric, SACF consistently outperforms SoundSpaces +3.6% on Heard and +2.0% on Unheard, but is slightly below AGSA on Replica 51.7% vs. 52.0% on Heard; 18.7% vs. 22.4% on Unheard. On Matterport3D [24], SACF yields improvements of +0.7%, +1.3%, and +1.2% in SPL, SR, and SNA on Heard scenes, and larger gains of +20.5%, +24.8%, and +17.6% on Unheard scenes. In particular, SACF attains 42.4%/58.3%/28.0% in SPL,SR, and SNA on the Unheard split, indicating strong generalization to unseen sound sources. We attribute these gains to modeling sound source locations as discrete distributions over direction and distance, which provides compact spatial priors from auditory cues to guide the policy while suppressing irrelevant visual regions. The slightly lower SNA on Replica, compared to AGSA, suggests a trade-off between navigation efficiency and final bearing alignment.

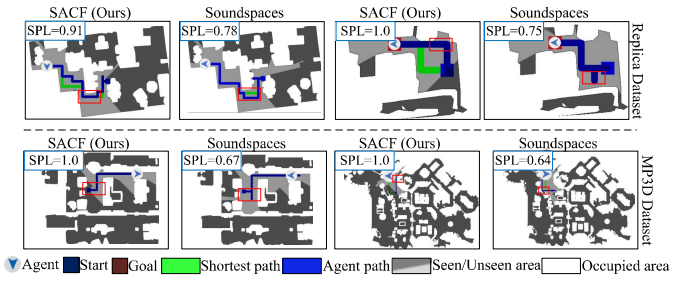

Fig. 4. Red boxes highlight key turning points where acoustic cues successfully guide SACF to avoid dead ends and navigate more efficiently than the baseline.

As shown in Table II, our SACF framework using RGB images as visual input also demonstrated improved navigation performance over the SoundSpaces[2] baseline. Since RGB images are susceptible to interference from lighting variations and texture details, depth maps directly provide geometric structural information about the environment. This enables agents to navigate obstacles more effectively and perceive spatial structures, making depth maps superior to RGB images for this task.

C. Ablation experiment

a) Module ablation: To validate the contributions of each SACF component, we conduct ablation studies on Replica and Matterport3D datasets across three configurations, Baseline: Both SDLD and ACVF are removed; audio and visual features from standard encoders are simply concatenated. w/o SDLD: ACVF is retained, but the FiLM layer is conditioned only on global audio embeddings without spatial descriptor guidance. w/o ACVF: SDLD is retained, but its spatial descriptor g_t is directly concatenated with unmodulated visual features, bypassing the channel modulation mechanism. As shown in Table III, incorporating either module independently outperforms the baseline, confirming their individual effectiveness.

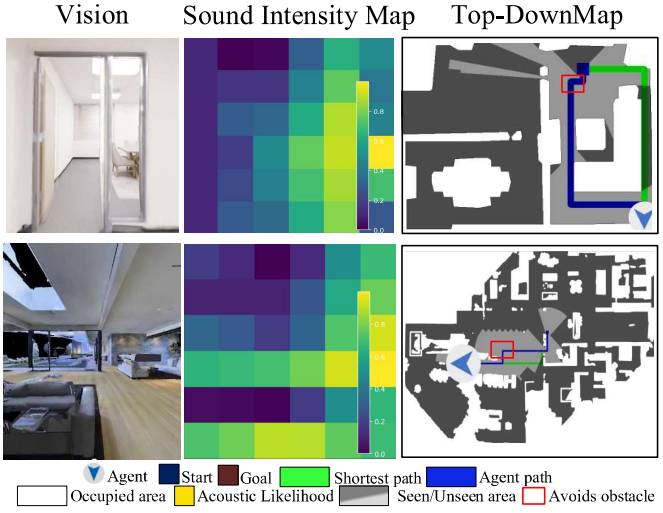

Fig. 5. Top-down navigation trajectory diagram for navigation.

b) Parameter Comparison: To demonstrate the lightweight advantages of ACVF, Table IV compares the efficiency of the three fusion schemes. Compared to Simple Concatenation, SACF adds only 0.15M parameters and is more compact than Cross-Modal Spatial Attention. Theoretically, unlike Cross-Modal Spatial Attention, which scales quadratically with spatial resolution, our FiLM-based ACVF operates with strictly linear computational complexity. Empirically, SACF maintains a high training throughput of approximately 48 FPS on Replica and approximately 74 FPS on Matterport3D. This ensures an extremely low real-time factor (RTF) during inference, confirming its lightweight nature and suitability for resource-constrained devices.

D. Qualitative analysis

To gain a more intuitive understanding of how our model operates, we conducted a qualitative analysis of several typical navigation trajectories. Fig. 4 and 5 illustrates the navigation trajectory on the top-down map of the SDLD model. The accompanying sound intensity map accurately pinpoints the

direction of the sound source, effectively guiding the agent to navigate toward the target. SACF uses a single fixed random seed for training and evaluation, as long-term reinforcement learning in a 3D environment requires significant computational resources. However, the reward and SPL curves shown in 6, converge smoothly and without oscillations, indicating that our proposed SACF framework exhibits stable learning dynamics. This also demonstrates that SACF achieves higher SPL values with fewer training steps, indicating that it converges faster than SoundSpaces.

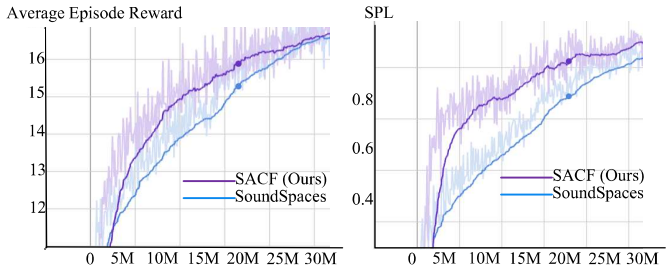

Fig. 6. Convergence: rewards and SPL (0–1; tables in %).

V. Conclusion

This paper proposes the Spatial-Aware Conditioned Fusion (SACF) framework, which effectively addresses the core challenges in existing audio-visual navigation methods, namely the ambiguous spatial intent of targets and the lack of directionality in visual representations, through explicit spatial discretization and deep cross-modal feature modulation. Experimental results demonstrate that the strategy combining Spatially Discretized Localization Descriptor (SDLD) with Audio-Descriptor Conditioned Visual Fusion (ACVF) not only significantly improves navigation efficiency on Replica and Matterport3D datasets, but also exhibits strong cross-scene generalization capabilities, particularly when encountering unfamiliar target sounds, while achieving an optimal balance between computational overhead and navigation performance through its lightweight FiLM-based design. This approach establishes a new effective paradigm for constructing efficient and robust embodied audio-visual navigation systems.

ACKNOWLEDGMENT

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

References

[1] 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.

[2] R. Gao, C. Chen, Z. Al-Halah, C. Schissler, and K. Grauman, "Visualechoes: Spatial image representation learning through echolocation," in *European Conference on Computer Vision*, 2020, pp. 658–676.

[3] 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.

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

[5] 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.

[6] 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.

[7] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[8] 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.

[9] 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.

[10] 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.

[11] C. Chen, S. Majumder, Z. Al-Halah, R. Gao, S. K. Ramakrishnan, and K. Grauman, "Learning to set waypoints for audio-visual navigation," in *Embodied Multimodal Learning Workshop at ICLR 2021*, 2021.

[12] 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.

[13] 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.

[14] 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.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] 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.

[17] 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.

[18] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," *Advances in neural information processing systems*, vol. 30, 2017.

[19] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," in *International Conference on Learning Representations (ICLR)*, 2020.

[20] H. Kondoh and A. Kanezaki, "Embodied navigation with auxiliary task of action description prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 7025–7036.

[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] H. Du, L. Li, Z. Huang, and X. Yu, "Object-goal visual navigation via effective exploration of relations among historical navigation states," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2563–2573.

[23] 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.

[24] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," in *2017 International Conference on 3D Vision (3DV)*, 2017, pp. 667–676.

[25] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9339–9347.