Generalizable Audio-Visual Navigation via Binaural Difference Attention and Action Transition Prediction

Jia Li[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*—In Audio-Visual Navigation (AVN), agents must locate sound sources in unseen 3D environments using visual and auditory cues. However, existing methods often struggle with generalization in unseen scenarios, as they tend to overfit to semantic sound features and specific training environments. To address these challenges, we propose the Binaural Difference Attention with Action Transition Prediction (BDATP) framework, which jointly optimizes perception and policy. Specifically, the Binaural Difference Attention (BDA) module explicitly models interaural differences to enhance spatial orientation, reducing reliance on semantic categories. Simultaneously, the Action Transition Prediction (ATP) task introduces an auxiliary action prediction objective as a regularization term, mitigating environment-specific overfitting. Extensive experiments on the Replica and Matterport3D datasets demonstrate that BDATP can be seamlessly integrated into various mainstream baselines, yielding consistent and significant performance gains. Notably, our framework achieves state-of-the-art Success Rates across most settings, with a remarkable absolute improvement of up to 21.6 percentage points in Replica dataset for unheard sounds. These results underscore BDATP's superior generalization capability and its robustness across diverse navigation architectures.**

***Index Terms*—Embodied AI, Audio-Visual Navigation, Reinforcement Learning, Binaural Difference Attention, Action Transition Prediction**

I. Introduction

With the rapid development of artificial intelligence [1]–[4] and robotics [5]–[7], autonomous navigation for embodied agents has received growing attention in both virtual and real-world environments [8]–[10]. Audio-Visual Navigation (AVN) is a representative multimodal task in which an agent is required to locate and reach a sound-emitting target using only egocentric visual observations and binaural audio signals in previously unseen environments [11]–[14]. Despite recent progress [15], [16], existing AVN methods still suffer from limited generalization ability [17]–[19]. Their performance often degrades significantly when evaluated on *unheard* sound categories or *unseen* spatial layouts. This limitation reveals fundamental challenges in learning robust multimodal representations and navigation policies that can generalize beyond training conditions [20], [21].

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

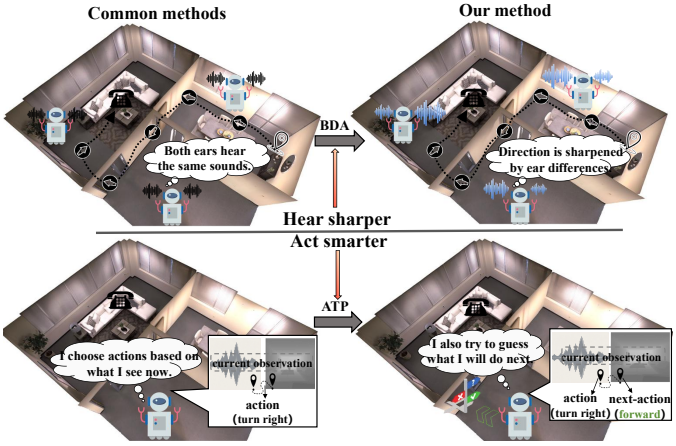

Fig. 1. Comparison between our BDATP and conventional AVN methods.

We identify two key factors underlying this issue. First, many audio representations implicitly entangle semantic content with spatial information [22]. While sound semantics may aid recognition, they are often unreliable under category shifts and can obscure universal localization cues [17], making directional inference unstable for unheard sounds. Second, reinforcement learning-based navigation policies are prone to overfitting to the dynamics and geometries of training environments [23], resulting in inefficient or unstable behaviors—such as oscillation or backtracking—when deployed in novel scenes.

To overcome these challenges, we propose a unified framework that enhances both perception and policy learning. From a perceptual perspective, the core idea is to prioritize binaural spatial differences over sound semantics, enabling the agent to *Hear Sharper* by focusing on directionally informative cues that generalize across sound categories. From a policy perspective, encouraging temporal consistency in action transitions regularizes navigation behavior, allowing the agent to *Act Smarter* with smoother and more stable trajectories in unseen environments. As illustrated in Fig. 1, this joint design fundamentally differs from conventional AVN pipelines by explicitly improving spatial perception and policy robustness

in a complementary manner.

Extensive experiments on the Replica [24] and Matterport3D [25] datasets demonstrate that the proposed framework achieves significant improvements over existing methods across comprehensive benchmarks, particularly in challenging zero-shot generalization settings. Our main contributions are summarized as follows:

- We propose a binaural spatial perception mechanism that emphasizes interaural differences, enabling agents to *Hear Sharper* under unheard sound categories.
- We introduce a policy regularization strategy that promotes temporally consistent action transitions, allowing agents to *Act Smarter* in unseen environments.
- We conduct comprehensive evaluations demonstrating strong generalization performance across novel sounds and unseen spatial layouts.

II. Related Work

Audio-Visual Navigation (AVN) has been extensively studied as a multimodal embodied task that integrates auditory and visual cues for target-driven navigation. Early work, such as SoundSpaces [11], established realistic acoustic simulation environments and demonstrated the feasibility of end-to-end reinforcement learning for sound-guided navigation. Subsequent approaches introduced architectural and algorithmic enhancements to improve navigation efficiency and robustness, including hierarchical planning with intermediate waypoints [16], semantic memory for intermittent sound sources [26], and omnidirectional perception for richer environmental awareness [27].

More recent generalization remains a central challenge in AVN, particularly when agents are evaluated on unheard sound categories or unseen environments. Several works attempt to reduce reliance on sound semantics by introducing auxiliary objectives or learning more abstract audio representations [17]. However, many existing methods still depend on implicit feature learning through black-box encoders, which may struggle to disentangle spatial cues from semantic content under distribution shifts. From the policy perspective, reinforcement learning-based navigation agents are known to overfit to the specific dynamics and layouts of training environments, leading to unstable trajectories in novel scenes [16], [27], [28]. While prior efforts have explored transfer learning or pre-training strategies to improve robustness, learning environment-agnostic and temporally consistent navigation behaviors remains an open problem in audio-visual navigation.

III. Method

To enhance generalization in AVN, we propose the BDATP framework (Fig. 2), comprising two core components: a **Binaural Difference Attention (BDA)** module to capture universal interaural spatial cues for robust perception, and an **Action Transition Prediction (ATP)** auxiliary task to reinforce cross-environment statistical regularities during policy learning.The following subsections detail these components.

A. Feature Extraction and Binaural Difference Attention

In the feature extraction stage, the visual input is a depth image, and the auditory input is a binaural spectrogram. Similar to AV-NaV, we employ two independent CNN encoders, each consisting of convolutional layers with kernel sizes of 8×8, 4×4, and 3×3, followed by a linear layer with ReLU activations in between. These encoders separately produce the visual feature $f_v \in \mathbb{R}^{512}$ and the audio feature $f_a \in \mathbb{R}^{512}$.

Conventional approaches often concatenate the binaural spectrograms directly, which tends to ignore interaural differences and weakens the agent's ability to perceive sound

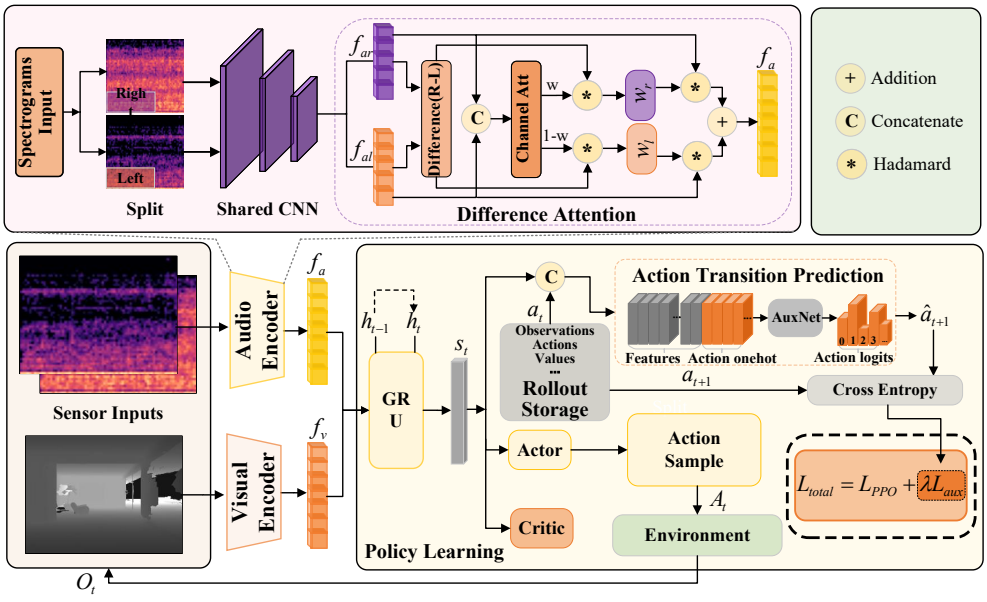

Fig. 2. BDATP framework overview: separate encoding of visual and auditory inputs, with pink (BDA) and yellow (ATP) modules.

TABLE I
PERFORMANCE COMPARISON ON REPLICA AND MATTERPORT3D DATASETS

Method	Replica						Matterport3D					
	Heard			Unheard			Heard			Unheard		
	SR↑	SPL↑	SNA↑	SR↑	SPL↑	SNA↑	SR↑	SPL↑	SNA↑	SR↑	SPL↑	SNA↑
Random Agent	18.5	4.9	1.8	18.5	4.9	1.8	9.1	2.1	0.8	9.1	2.1	0.8
Direction Follower	72.0	54.7	41.1	17.2	11.1	8.4	41.2	32.3	23.8	18.0	13.9	10.7
SAVi [26]	54.0	45.1	30.8	33.9	27.5	17.2	40.3	29.1	13.0	29.5	20.4	9.6
Dav-NaV [29]	85.1	72.6	54.0	58.5	45.6	33.4	82.9	61.9	46.8	55.3	42.4	31.6
SA2GVAN [17]	90.4	70.9	55.2	62.8	43.4	33.0	82.9	61.4	46.8	60.7	42.3	31.4
ORAN [27]	-	-	-	60.9	46.7	**36.5**	-	-	-	59.4	**50.8**	**35.2**
AV-NaV [11]	88.9	64.5	44.1	47.3	34.7	14.1	66.2	44.8	27.3	33.5	21.9	10.4
AV-NaV + BDATP	93.1	74.5	43.9	68.6	45.0	19.4	68.7	51.7	28.2	55.1	37.9	20.1
AV-WaN [16]	90.9	70.4	52.5	52.8	34.7	27.1	82.4	55.4	42.5	56.7	40.9	30.4
AV-WaN + BDATP	**96.5**	**79.2**	**63.5**	**70.7**	**49.9**	34.6	**85.4**	**66.4**	**52.1**	**65.4**	44.0	32.7

direction, especially leading to degraded performance on unseen sound categories [30]. To address this, we introduce the BDA module within the audio encoder, which is applied before the linear layer. Specifically, the binaural spectrogram is split into the left and right channels before encoding. A CNN with shared weights encodes them separately, yielding intermediate feature maps f_{al} and f_{ar}. We then compute the element-wise difference and concatenate the original channel features:

$$diff = |f_{ar} - f_{al}| \in \mathbb{R}^{B\times C\times h\times w}, \tag{1}$$

$$f'_a = \mathrm{Concat}(f_{al}, f_{ar}) \in \mathbb{R}^{B\times 2C\times h\times w}. \tag{2}$$

The concatenated feature f'_a is projected back from $2C$ channels to C via a 1×1 convolution, followed by a Sigmoid activation to obtain channel-wise attention weights:$w \in (0,1)^{B\times C\times h\times w}$. Unlike conventional interpolation-based weighting, we explicitly use $diff$ as a modulation term, with w distributing emphasis between the left and right channels:

$$w_r = w \odot diff, \tag{3}$$

$$w_l = (1 - w) \odot diff, \tag{4}$$

$$f_a = f_{al} \odot w_l \ + \ f_{ar} \odot w_r. \tag{5}$$

Here, w_r and w_l are the difference-weighted coefficients for the right and left channels, and f_a is the fused audio feature.Compared to simple concatenation or weighting, BDA enforces attention to binaural spatial differences while preserving original channel amplitudes. This allows the model to leverage cross-category spatial cues for better generalization to unseen sounds.

B. Policy Learning with Action Transition Prediction

In reinforcement learning-based AVN, sparse reward signals often cause policies to overfit to training environments, resulting in unstable trajectories and poor generalization to unseen scenes. To mitigate this, we propose the Action Transition Prediction (ATP) auxiliary task. ATP leverages intra-rollout temporal supervision to impose a global statistical regularization across parallel environments. Specifically, it penalizes the discrepancy between the predicted action for the next timestep and the actual action executed by the agent within the same trajectory. By incorporating this prediction error as a regularization term, ATP encourages the policy to prioritize environment-invariant features—extracting cross-modal representations essential for navigation.

Formally, at each timestep t, given the current state feature s_t and the action a_t taken at the previous timestep, an auxiliary network predicts the next action $\hat{a}_{t+1}$:

$$\hat{a}_{t+1} = \mathrm{AuxNet}([s_t; \mathrm{OneHot}(a_t)]), \tag{6}$$

where AuxNet is a two-layer fully-connected network. The cross-entropy loss is computed using the ground-truth action a_{t+1} that was actually performed at the next timestep within the same rollout:

$$\mathcal{L}_{aux} = \frac{1}{(T-1)\cdot B} \sum_{i=1}^{(T-1)\cdot B} -\log \frac{\exp([\hat{a}_{t+1}]_{i,(a_{t+1})_i})}{\sum_{j=1}^{N} \exp([\hat{a}_{t+1}]_{i,j})}, \tag{7}$$

where $[\hat{a}_{t+1}]_{i,j}$ represents the predicted logits for sample i and action j, $(a_{t+1})_i \in \{0, \ldots, N-1\}$ is the ground-truth action index, B is the number of parallel environments (batch size), and T is the rollout length. Specifically,the value of N denotes the size of the action space, which varies depending on the navigation backbone:For AV-NaV, $N = 4$ corresponding to discrete low-level actions (Forward, Left, Right, Stop).For AV-WaN, $N = 81$ as the model predicts intermediate waypoints across a 9×9 spatial action map.

Finally, within the PPO framework, the auxiliary loss is added as a regularization term to the policy loss:

$$\mathcal{L}_{total} = \mathcal{L}_{PPO} + \lambda \mathcal{L}_{aux}, \tag{8}$$

where $\mathcal{L}_{PPO}$ includes the clipped surrogate loss, value function loss, and entropy regularization. We set $\lambda = 0.1$ to balance auxiliary supervision with reward-driven optimization.

The key insight of ATP is that while specific states s_t vary across environments, the action transitions reflect consistent navigational regularities. For instance, agents generally exhibit similar turning probabilities when encountering obstacles or lateral sound gradients regardless of the specific room geometry. Although parallel rollouts yield diverse state-action pairs,

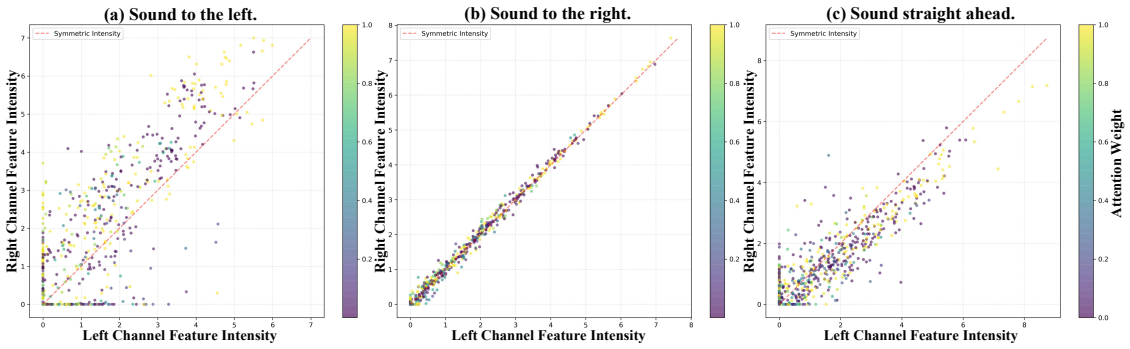

Fig. 3. Feature relationship visualization of the BDA module across three spatial scenarios.

they all encode these common underlying transition dynamics. By jointly optimizing the transition prediction loss across a batch of rollouts, ATP enforces a shared statistical constraint, strengthening the policy's sensitivity to generalizable cues and improving transferability to novel environments.

IV. Experiments and Analysis

To assess the effectiveness and generalization of our method in audiovisual navigation, we conduct experiments on two realistic 3D simulation datasets and compare with state-of-the-art baselines. These baselines include heuristic methods (**Random** and **Direction Follower**), the fundamental RL-based **AV-NaV** [11], the hierarchical waypoint-based **AV-WaN** [16], the semantic-aware **SAVi** [26], the robust spatial-fusing **Dav-NaV** [29], the omnidirectional navigator **ORAN** [27], and the semantic-agnostic **SA2GVAN** [17]. This section details the experimental setup, quantitative and qualitative results, and trajectory analysis.

A. Experimental Setup

Experiments are conducted in the SoundSpaces [11] simulation platform using two real-world 3D scene datasets: Replica [24] and Matterport3D [24]. The Replica environments are relatively small (average area 47.24 m^2), while Matterport3D contains larger and more complex environments (average area 517.34 m^2). To demonstrate the effectiveness of our framework, we validate our method on two representative baselines, AV-NaV [11] and AV-WaN [16].

Navigation performance is evaluated by the following metrics [31]: (1) **SR** (Success Rate): measures whether the agent successfully reaches the target location. (2) **SPL** (Success weighted by Path Length): combines success rate with path efficiency. (3) **SNA** (Success weighted by Number of Actions): assesses the economy of agent behavior by weighing success against the ratio of optimal to actual execution steps.

We consider two evaluation settings based on sound category exposure: **Heard** and **Unheard**, both involving multiple sound types. In the Heard setting, test sound categories appear in training but test scenes are unseen. In the Unheard setting, both sound categories and scenes are unseen.

B. Results and Analysis

1) Quantitative Results: Table I reports the performance comparison on Replica and Matterport3D datasets. Overall, our BDATP framework consistently enhances the navigation capabilities of mainstream baselines, achieving state-of-the-art results across most metrics.

Specifically, on the Replica dataset, integrating BDATP with AV-WaN yields the highest performance. In the *Heard* setting, BDATP improves the SR of AV-WaN from 90.9% to 96.5% and SPL from 70.4% to 79.2%. The advantage is even more pronounced in the challenging *Unheard* setting, where AV-WaN + BDATP achieves an SR of 70.7% and an SPL of 49.9%, outperforming the original AV-WaN baseline by 17.9% and 15.2%, respectively. Similarly, when applied to the AV-NaV baseline, BDATP boosts the *Unheard* SR from 47.3% to 68.6%, a substantial improvement of 21.3 percentage points. These results validate that the BDA module effectively captures essential binaural spatial cues, while the ATP task successfully regularizes the policy for better generalization to novel sound categories.

On the larger and more acoustically complex Matterport3D dataset, BDATP maintains robust performance gains. The AV-WaN + BDATP configuration achieves the best SR of 85.4% in the *Heard* setting and 65.4% in the *Unheard* setting. Notably, even when integrated with the simpler AV-NaV architecture, BDATP boosts the *Unheard* SR by 21.6 percentage points (from 33.5% to 55.1%). This consistent improvement across different architectures and dataset scales underscores the plug-and-play nature of our framework. Furthermore, the significant increase in SNA across most experimental settings confirms that our method not only reaches the goal more frequently but also does so with fewer redundant actions, resulting in more efficient and stable navigation trajectories.

2) Qualitative Analysis: To better understand BDATP's generalization ability, we analyze binaural perception, action transitions, and navigation trajectories. As shown in Fig. 3, binaural feature distributions under directional acoustic conditions deviate from the diagonal ($y = x$), indicating strong sensitivity to interaural differences and enabling robust spatial localization even for unheard sound categories. Fig. 4 shows that the ATP auxiliary task produces a clear diagonal pattern

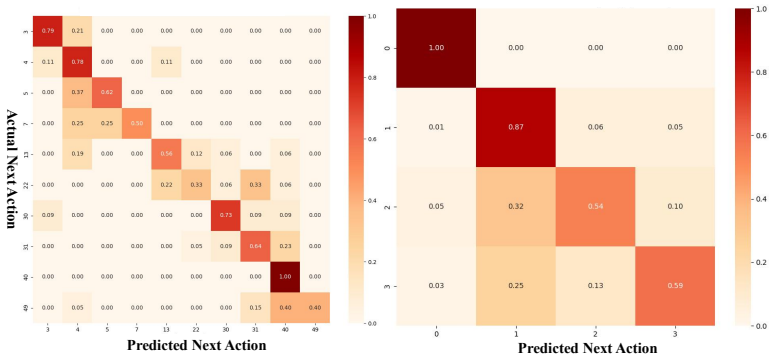

Fig. 4. Action Transition Matrices of ATP.Left: Top-10 most probable transitions on AV-WaN; Right: Full transitions on AV-Nav.

TABLE II
ABLATION STUDY RESULTS ON REPLICA DATASETS

Model	Heard		Unheard	
	SR↑	SPL↑	SR↑	SPL↑
w/o BDA and ATP	88.9	64.5	47.3	34.7
w/o ATP	90.2	74.0	66.2	44.4
w/o BDA	92.2	72.9	63.4	44.3
AV-NaV + BDATP	**93.1**	**74.5**	**68.6**	**45.0**

TABLE III
ABLATION STUDY OF THE AUXILIARY TASK LOSS WEIGHT λ ON REPLICA

Weight λ	Heard		Unheard	
	SR↑	SPL↑	SR↑	SPL↑
$\lambda = 0$	88.9	64.5	47.3	34.7
$\lambda = 0.001$	90.2	66.3	57.1	39.2
$\lambda = 0.01$	88.9	68.8	62.9	39.1
$\lambda = 0.1$(Ours)	**92.2**	**72.9**	**63.4**	**44.3**

in action transition matrices, reflecting high consistency with ground truth and reduced transition ambiguity. As illustrated in Fig. 5, BDATP generates smoother and more stable trajectories on both Matterport3D and Replica datasets in the Unheard setting, while baseline methods often fail or exhibit inefficient behaviors. These results demonstrate that enhancing spatial perception and action transition reliability leads to more efficient and robust navigation toward novel sound sources.

3) Ablation Study: We evaluate the contributions of BDA and ATP on the AV-NaV architecture using the Replica dataset (Table II). Removing both modules causes a clear performance drop, especially in the *Unheard* setting, where SR falls to 47.3%, highlighting the need for both spatial perception and policy learning. Without BDA, the *Unheard* SR decreases from 68.6% to 63.4%, showing that modeling interaural differences is crucial for generalizing to novel sound categories. Removing

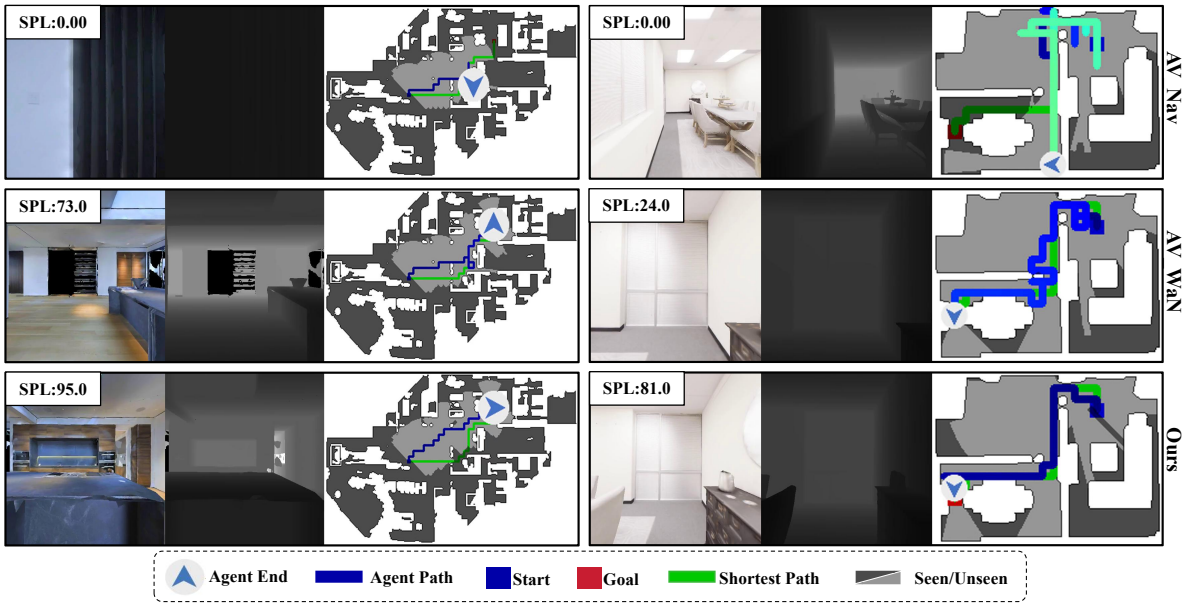

Fig. 5. Trajectory comparison in unseen/unheard settings: Matterport3D (left) and Replica (right).

ATP mainly affects stability and efficiency: while *Heard* performance remains strong, *Unheard* SR drops to 66.2% and SPL is consistently lower. Overall, BDA provides reliable directional cues, while ATP improves policy robustness through more consistent action transitions.

We further analyze the sensitivity to the auxiliary task loss weight λ based on the AV-NaV architecture without BDA (Table III), where $\lambda = 0$ represents the vanilla AV-NaV baseline.Increasing λ to 0.1 consistently enhances performance, particularly boosting the Unheard SR from 47.3% to 63.4%. This 16.1 percentage point improvement confirms that the ATP task provides crucial policy regularization.

V. Conclusion

This paper presents **BDATP**, a framework designed to enhance generalization in Audio-Visual Navigation. By integrating **BDA** for robust spatial perception and **ATP** for shared statistical regularization of action transitions, our method effectively mitigates overfitting to specific sound categories and environments. Extensive evaluations on Replica and Matterport3D confirm that BDATP serves as a versatile plug-and-play method, consistently improving navigation performance across multiple baselines. Future work will extend this framework to dynamic sound sources and multi-agent coordination in real-world acoustic environments.

ACKNOWLEDGMENT

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

References

[1] 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.

[2] Y. Yu, L. Cao, F. Sun, C. Yang, H. Lai, and W. Huang, "Echo-enhanced embodied visual navigation," *Neural Computation*, vol. 35, pp. 958–976, 2023.

[3] N. Jaquier, M. C. Welle, A. Gams, K. Yao, B. Fichera, A. Billard, A. Ude, T. Asfour, and D. Kragic, "Transfer learning in robotics: An upcoming breakthrough? a review of promises and challenges," *The International Journal of Robotics Research*, vol. 44, pp. 465–485, 2025.

[4] 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.

[5] 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.

[6] Z. Durante, Q. Huang, N. Wake, R. Gong, J. S. Park, B. Sarkar, R. Taori, Y. Noda, D. Terzopoulos, Y. Choi *et al.*, "Agent ai: Surveying the horizons of multimodal interaction," *CoRR*, 2024.

[7] W. Wu, T. Chang, X. Li, Q. Yin, and Y. Hu, "Vision-language navigation: a survey and taxonomy," *Neural Computing and Applications*, vol. 36, no. 7, pp. 3291–3316, 2024.

[8] C. Wen, Y. Huang, H. Huang, Y. Huang, S. Yuan, Y. Hao, H. Lin, Y.-S. Liu, and Y. Fang, "Zero-shot object navigation with vision-language models reasoning," in *International Conference on Pattern Recognition*, 2025, pp. 389–404.

[9] M. Iftikhar, M. Saqib, M. Zareen, and H. Mumtaz, "Artificial intelligence: revolutionizing robotic surgery," *Annals of Medicine and Surgery*, pp. 5401–5409, 2024.

[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] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 2020, pp. 17–36.

[12] 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.

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

[15] 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, 2025, pp. 14 673–14 680.

[16] C. Chen, S. Majumder, A.-H. Ziad, R. Gao, S. Kumar Ramakrishnan, and K. Grauman, "Learning to set waypoints for audio-visual navigation," in *ICLR*, 2021.

[17] H. Wang, Y. Wang, F. Zhong, M. Wu, J. Zhang, Y. Wang, and H. Dong, "Learning semantic-agnostic and spatial-aware representation for generalizable visual-audio navigation," *IEEE Robotics and Automation Letters*, vol. 8, pp. 3900–3907, 2023.

[18] 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.

[19] 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.

[20] Y. Wu, P. Zhang, M. Gu, J. Zheng, and X. Bai, "Embodied navigation with multi-modal information: A survey from tasks to methodology," *Information Fusion*, p. 102532, 2024.

[21] Z. Zhao, H. Tang, and Y. Yan, "Audio-visual navigation with anti-backtracking," in *International Conference on Pattern Recognition*, 2025, pp. 358–372.

[22] 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.

[23] A. S. Roman, A. Chang, G. Meza, and I. R. Roman, "Generating diverse audio-visual 360 soundscapes for sound event localization and detection," *arXiv e-prints*, pp. arXiv–2504, 2025.

[24] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green *et al.*, "The Replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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, pp. 928–935, 2023.

[30] 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.

[31] 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.