Reliability-Aware Geometric Fusion for Robust Audio-Visual Navigation

Teng Liu[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 (AVN) requires an embodied agent to move toward a sounding target using both visual observations and binaural audio signals. However, in complex indoor environments, binaural cues are not always reliable. This problem becomes more noticeable when the agent encounters sound categories that were not seen during training. In this work, we present RAVN (Reliability-Aware Audio-Visual Navigation), a framework that adjusts cross-modal fusion according to the reliability of audio observations. Instead of assuming that audio information is always trustworthy, our method estimates its reliability and uses this estimate to guide how audio and visual features are combined. To achieve this, we design an Acoustic Geometry Reasoner (AGR) trained with geometric proxy supervision. AGR learns observation-dependent uncertainty through a heteroscedastic Gaussian NLL objective. The learned uncertainty serves as a practical reliability signal and does not require geometric labels at inference time. Leveraging this signal, our Reliability-Aware Geometric Modulation (RAGM) module applies a soft gate to the incoming visual features. This explicitly suppresses inter-modal conflicts at the fusion stage. To test the architecture, we benchmarked RAVN within the SoundSpaces framework across Replica and Matterport3D layouts. Our comprehensive evaluations show consistent improvements over previous baselines, both in heard and unheard scenarios.**

***Index Terms*—Embodied AI, Audio-Visual Navigation, Multimodal Fusion, Uncertainty Estimation**

I. Introduction

Embodied navigation involves robotic agents (with egocentric observation) exploring the unknown environment to reach target locations [1]. But acoustic tracking is rarely straightforward. Room geometry, physical barriers, and sound wave diffraction constantly corrupt the incoming audio. Because of this severe physical distortion, standard binaural cues can quickly shift from being helpful to highly unreliable. As a result, precise localization in chaotic acoustic environments is fundamentally limited. The challenge grows when agents must adapt to unfamiliar sound sources. Unheard sounds introduce a shift in audio features, making it harder to extract reliable binaural cues and increasing the risk of following incorrect patterns or drifting off course [2]–[4].

This motivates the need for a critical yet underexplored capability in AVN: reliability-aware geometric fusion—dynamically modulating cross-modal integration based on inferred geometric reliability. This approach fully leverages audio information when cues are consistent. It reduces the influence of audio when the evidence becomes ambiguous. Consider human intuition: when hearing a distant or reverberant sound, we don't blindly follow it. Instead, we instinctively assess the sound's reliability before deciding how much to trust it. If the sound is unclear, we reduce the impact of auditory cues and prioritize visual confirmation [5].

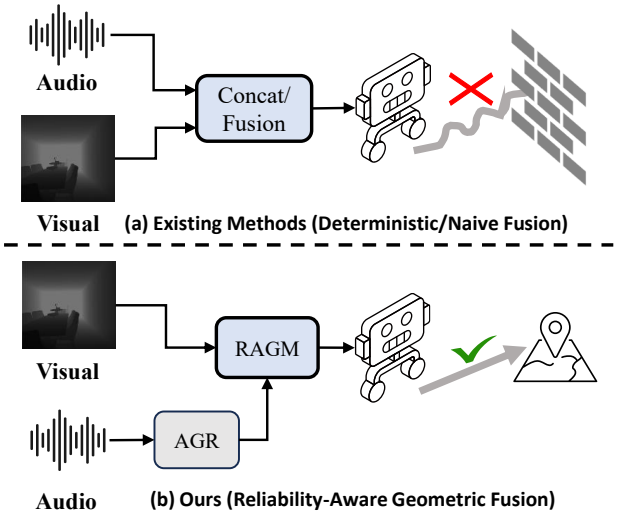

Fig. 1. **Paradigm comparison.** (a) naive deterministic fusion vs. (b) our reliability-aware RAVN framework.

Most existing AVN methods lack such mechanisms. First, they implicitly assume a high degree of signal reliability, prioritizing precise geometric regression while underestimating the stochastic nature of distorted audio, and rely on naive concatenation for modality fusion, making policies vulnerable to acoustic illusions in distorted environments [6]–[8]. Second, while auxiliary tasks are sometimes used for representation learning, their predictions are generally limited to training objectives and don't directly inform fusion at test time. As a result, agents struggle to adaptively calibrate their trust in audio under varying environmental conditions.

To address this gap and inspired by human intuition, we propose Reliability-Aware Audio-Visual Navigation (RAVN), as conceptually contrasted with existing methods in Fig. 1. Unlike static reliability assumptions, RAVN learns reliability-

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

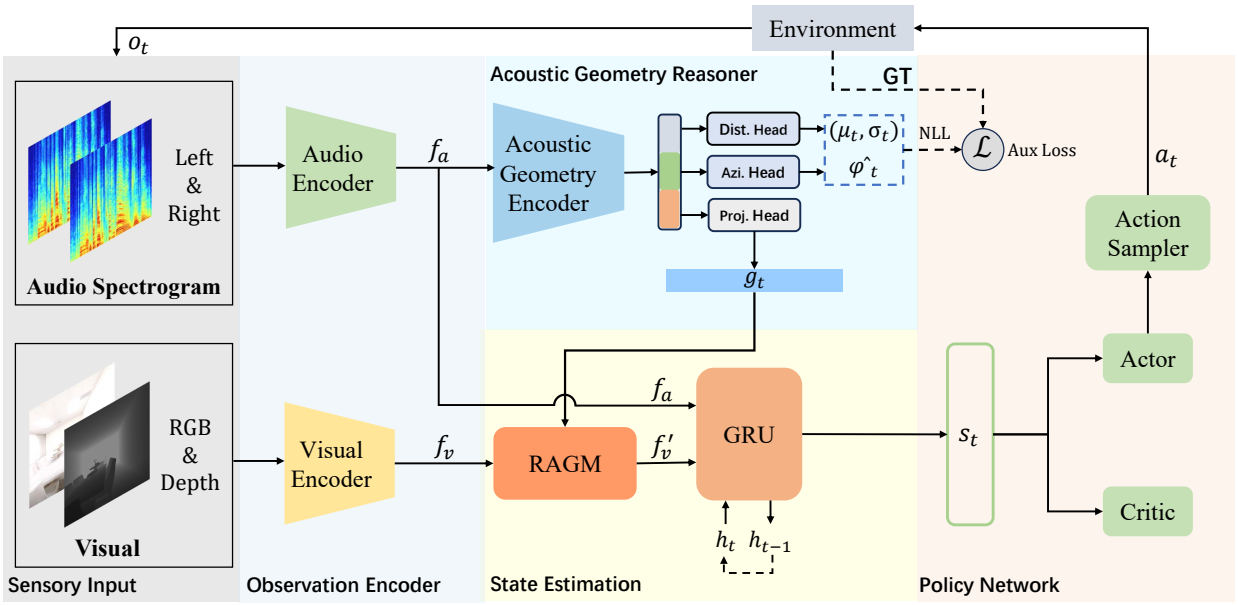

Fig. 2. **The RAVN Framework.** Multi-modal observations o_t are encoded into f_a and f_v. The AGR module estimates geometric embeddings g_t and predictive uncertainty (μ_t, σ_t), supervised by Ground Truth (GT) labels via an auxiliary loss. The RAGM module then uses g_t to dynamically modulate visual features into f_v'. A recurrent policy (GRU) aggregates these components into the hidden state s_t for end-to-end action (a_t) selection.

aware cues from binaural observations to adjust visual features. It filters out unreliable auditory guidance and reduces cross-modal conflicts. It lowers the weight of misleading cues and uses consistent evidence to adjust the fusion process. This helps with navigation in complex acoustic environments.

In summary, we present the following contributions:

- We introduce RAVN, a framework that learns reliability cues from audio using geometric supervision and a heteroscedastic objective.
- We introduce RAGM, a novel mechanism that utilizes these learned cues to dynamically modulate visual features, adjusting them according to audio reliability, thereby mitigating conflicts between modalities.
- We show with results that modeling sensory reliability helps improve navigation, especially in difficult scenarios with unheard sound sources.

II. Related Work

Embodied visual navigation mainly uses reinforcement learning in simulators like Habitat [9]. It has limits due to small fields of view. This leads to Audio-Visual Navigation (AVN), which uses binaural cues to help agents find sound-emitting targets [1], [10]. Existing AVN methods typically fuse modalities via static mechanisms like concatenation or attention [11]–[13]. However, these approaches implicitly assume constant auditory reliability. In complex acoustics, factors like reverberation and multipath propagation heavily distort binaural cues [14]. Consequently, static fusion leaves agents vulnerable to over-trusting misleading audio, especially under distribution shifts from unheard sounds. To address such perceptual ambiguity, uncertainty modeling—such as heteroscedastic regression—is widely used in general reinforcement learning [15], [16]. Yet, its direct integration into AVN fusion remains underexplored. While some AVN frameworks employ geometric auxiliary tasks for representation learning [17], [18], they aim to enhance deterministic feature distinctiveness rather than quantify signal quality. Similarly, recent probabilistic AVN methods [19], [20] heavily rely on precisely calibrated uncertainty, which often proves brittle under realistic acoustic shifts. Our work bridges this gap by explicitly learning audio-derived reliability to dynamically condition cross-modal fusion. Instead of requiring fragile, fully calibrated probabilistic inference, RAVN repurposes geometric auxiliary tasks to capture heteroscedastic uncertainty (i.e., variance in distance and azimuth inference) as a latent discriminative signal. This allows the agent to adaptively down-weight ambiguous auditory cues and fully leverage consistent evidence, ensuring robust navigation in chaotic acoustic environments.

III. The Proposed Approach

A. Overview

We formalize the problem as a reinforcement learning task where the agent learns a policy to efficiently navigate toward an active acoustic source in an unknown environment. At each time step t, the agent receives a raw multi-modal observation o_t comprising visual depth maps and binaural audio. To navigate robustly under complex acoustics, we propose RAVN

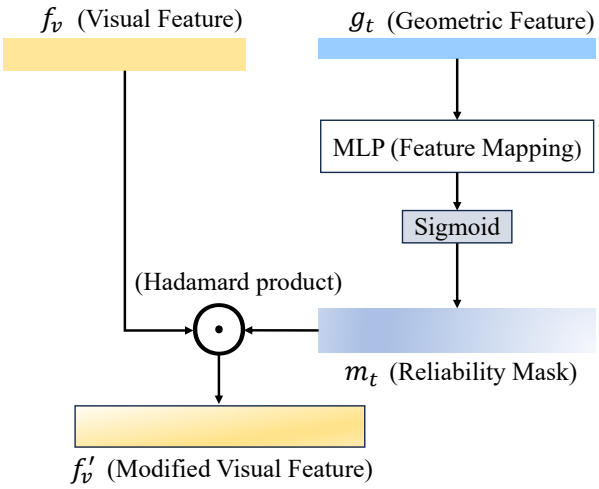

Fig. 3. **RAGM module architecture.** Geometric features are transformed into a reliability mask, which applies element-wise modulation to visual features to produce modified visual representations.

(Reliability-Aware Audio-Visual Navigation), a framework that leverages explicitly estimated acoustic reliability to guide cross-modal fusion. As shown in Fig. 2, RAVN consists of three core components: (i) an Acoustic Geometry Reasoner (AGR) that extracts ambiguity-aware geometric embeddings and heteroscedastic uncertainty from audio features; (ii) a Reliability-Aware Geometric Modulation (RAGM) module that dynamically recalibrates visual features via soft-gating, conditioned on the learned reliability; and (iii) a recurrent policy network that aggregates the modulated features to update its hidden state representation s_t for end-to-end action selection.

B. Acoustic Geometry Reasoner (AGR)

The Acoustic Geometry Reasoner (AGR) comprises a shared Acoustic Geometry Encoder and three parallel branches: a primary geometric projection head for feature alignment, and two auxiliary heads for representation learning. The encoder processes the CNN-extracted audio features f_a to extract spatial geometric information (such as relative distance and azimuth) and reliability data, producing a latent geometric embedding z_{geo}. Then, the projection head maps z_{geo} to a geometric representation g_t, aligning the acoustic features with the visual features to allow for element-wise modulation.

Our main aim is not precise localization, but rather learning representations that capture how dispersion depends on the observations. Instead of focusing on full calibration, the heteroscedastic objective provides useful reliability cues, which help adjust the fusion process for more robust navigation.

Probabilistic Distance Head: We model the distance as a Gaussian distribution, $d_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$, and minimize the Negative Log-Likelihood (NLL):

$$\mathcal{L}_{dist} = \frac{1}{2}\log\sigma_t^2 + \frac{(y_{dist} - \mu_t)^2}{2\sigma_t^2}, \tag{1}$$

where y_{dist} denotes the true geodesic distance. The variance, σ_t^2, captures aleatoric uncertainty and serves as a dynamic indicator of acoustic reliability. By minimizing NLL, we encourage σ_t^2 to increase when the prediction error is large, signaling a drop in reliability.

Azimuth Head: Along with the distance, we also predict the relative azimuth $\hat{\varphi}_t$ to enable orientation-aware feature learning. To address angular periodicity (i.e., the discontinuity between π and $-\pi$), we minimize the wrapped Smooth-L1 loss:

$$\mathcal{L}_{ang} = \text{SmoothL1}\left(\mathcal{W}(\hat{\varphi}_t - y_{ang})\right), \tag{2}$$

where y_{ang} is the true azimuth, and $\mathcal{W}(\cdot)$ wraps the error to stay within the $[-\pi, \pi]$ range. This approach ensures that the final embedding g_t captures crucial directional information, providing the spatial context necessary to complement the distance-based reliability.

Projection Head: We map the latent embedding z_{geo} to a geometric representation $g_t \in \mathbb{R}^{D_v}$ using a learnable function ϕ_{proj}:

$$g_t = \phi_{proj}(z_{geo}). \tag{3}$$

This projection serves two purposes: (1) it aligns the acoustic features with the visual features, allowing for element-wise fusion, and (2) guided by uncertainty gradients, g_t captures both spatial direction and acoustic reliability, acting as a compact representation for geometric reasoning with reliability awareness.

C. Reliability-Aware Geometric Modulation (RAGM)

The RAGM module converts the learned reliability cues into a fusion mechanism. As shown in Fig. 3, it takes the CNN-extracted visual feature f_v and the geometric feature g_t as inputs. We create a soft modulation mask $m_t \in (0,1)^{D_v}$ by passing g_t through an MLP and a sigmoid activation:

$$m_t = \text{Sigmoid}(\text{MLP}(g_t)). \tag{4}$$

Next, the visual feature is adjusted through an element-wise Hadamard product:

$$f_v' = f_v \odot m_t. \tag{5}$$

The reliability mask m_t works as a soft gate, dynamically changing based on the acoustic context. Since g_t shares its encoder with the probabilistic heads, it captures noise cues through joint optimization. During navigation, m_t dynamically balances the fusion. If the sound is bad, it drops the cross-modal features that could lead to wrong turns. When the audio is reliable, it boosts the visual signals to quickly locate the target. This element-wise modulation allows the fusion process to remain flexible, enabling the agent to adapt its use of visual and auditory information based on the environment.

D. Policy Learning via State Estimation

We concatenate the raw audio and the modulated visual features as $x_t = [f_a, f_v']$ and feed them into a GRU:

$$s_t = \text{GRU}(x_t, h_{t-1}), \tag{6}$$

TABLE I

QUANTITATIVE COMPARISON ON REPLICA AND MATTERPORT3D DATASETS. WE REPORT SUCCESS RATE (SR), SUCCESS WEIGHTED BY PATH LENGTH (SPL), AND SUCCESS WEIGHTED BY NUMBER OF ACTIONS (SNA). BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Replica						Matterport3D					
	Heard			Unheard			Heard			Unheard		
	SR ↑	SPL ↑	SNA ↑	SR ↑	SPL ↑	SNA ↑	SR ↑	SPL ↑	SNA ↑	SR ↑	SPL ↑	SNA ↑
Random Agent [17]	18.5	4.9	1.8	18.5	4.9	1.8	9.1	2.1	0.8	9.1	2.1	0.8
Frontier Waypoints [17]	63.9	44.0	35.2	14.8	6.5	5.1	42.8	30.6	22.2	16.4	10.9	8.1
Direction Follower [17]	72.0	54.7	41.1	17.2	11.1	8.4	41.2	32.3	23.8	18.0	13.9	10.7
Supervised Waypoints [17]	88.1	59.1	48.5	43.1	14.1	10.1	36.2	21.0	16.2	8.8	4.1	2.9
Gan et al. [19]	83.1	57.6	47.9	15.7	7.5	5.7	37.9	22.8	17.1	10.2	5.0	3.6
AV-Nav [1]	93.0	76.1	44.7	47.3	36.8	21.5	68.8	52.3	29.6	34.5	25.0	12.7
RAVN (Ours)	**97.0**	**80.2**	**52.4**	**53.1**	**40.3**	**23.4**	**70.9**	**55.6**	**33.2**	**35.6**	**25.9**	**13.6**

where h_t is the recurrent hidden state and s_t is the state representation used by the actor and critic to produce $\pi(a_t|s_t)$ and $V(s_t)$.

Importantly, the modulated visual feature f'_v is concatenated with the unmodulated audio feature f_a before entering the GRU. While f'_v adaptively gates visual exploratory cues based on environmental reliability, preserving the raw f_a ensures the policy continuously receives the primary goal-directed auditory signal.

E. Training Objectives

The entire framework is trained end-to-end. While the primary policy is optimized via PPO [21] to maximize navigation rewards, the robust representation learning in AGR is driven by Auxiliary Geometric Supervision ($\mathcal{L}_{aux}$).

Formally, we aggregate the supervision signals from the two auxiliary heads defined in Sec. III-B:

$$\mathcal{L}_{aux} = \mathcal{L}_{dist} + \mathcal{L}_{ang}, \tag{7}$$

where $\mathcal{L}_{dist}$ is the heteroscedastic regression loss that instills uncertainty awareness, and $\mathcal{L}_{ang}$ denotes the cross-entropy loss for direction estimation. The overall training objective is a weighted combination:

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

where λ balances the reinforcement learning objective with geometric representation learning.

IV. EXPERIMENTS

A. Experimental Setup

We implement our framework on the SoundSpaces platform using two publicly available datasets: Replica, which contains 18 scanned apartments and offices with 0.5m grids, and Matterport3D, comprising 85 large-scale home environments with 1m grids [1], [22], [23]. The RIR sampling rates are set to 44.1 kHz for Replica and 16 kHz for Matterport3D. We train the model using the Adam optimizer with a learning rate of 2.5×10^{-4}. We train for 48 million steps on Replica and 60 million steps on Matterport3D, with each episode limited to 500 steps. Following standard evaluation protocols, we consider two settings: (1) Multiple Heard, where all 102 sounds are included in the training, validation, and test splits; and (2) Multiple Unheard, where the sounds are divided into non-overlapping splits of 73/11/18 to test generalization.

B. Evaluation Metrics

Following standard protocols, we use three metrics to assess navigation performance: **Success Rate (SR)**, which measures the percentage of episodes where the agent successfully reaches the goal within the time limit; **Success weighted by Path Length (SPL)**, our primary efficiency metric, which factors in the ratio of the shortest-path (geodesic) distance to the actual path length, reflecting the agent's ability to avoid unnecessary detours; and **Success weighted by Number of Actions (SNA)**, which penalizes excessive actions and complements SPL by capturing decisiveness and stability, especially under sensory ambiguity.

C. Quantitative Evaluation

Task configurations and evaluation setups strictly follow the SoundSpaces rules. All metric breakdowns are available in Table I. Looking at Replica, the AV-Nav baseline simply fails to match RAVN under any sound condition. Take the Unheard task as the primary example. Here, SR improved by 12.3% in relative terms. We also measured a 9.5% relative gain for SPL. Hitting a 53.1% success rate—along with taking much more direct paths—means the agent adapts stably to new acoustic scenarios. When encountering familiar audio (the Heard setting), the agent succeeds almost every time, resting at a 97.0% SR. SNA also went up by 17.2%. The agent moves with practically zero hesitation.

Matterport3D presents much harder spatial layouts, yet the performance gap holds. For Heard sounds, we recorded a 70.9% SR, and SNA improved by 12.2% relative to the baseline. The most severe test is the Unheard setting on this dataset. Finding completely new sounding targets is notoriously tough. Still, our method managed to secure a 35.6% success rate. That gives RAVN a 3.2% margin over the baseline. This validates that reliability-aware fusion remains robust in complex spatial layouts.

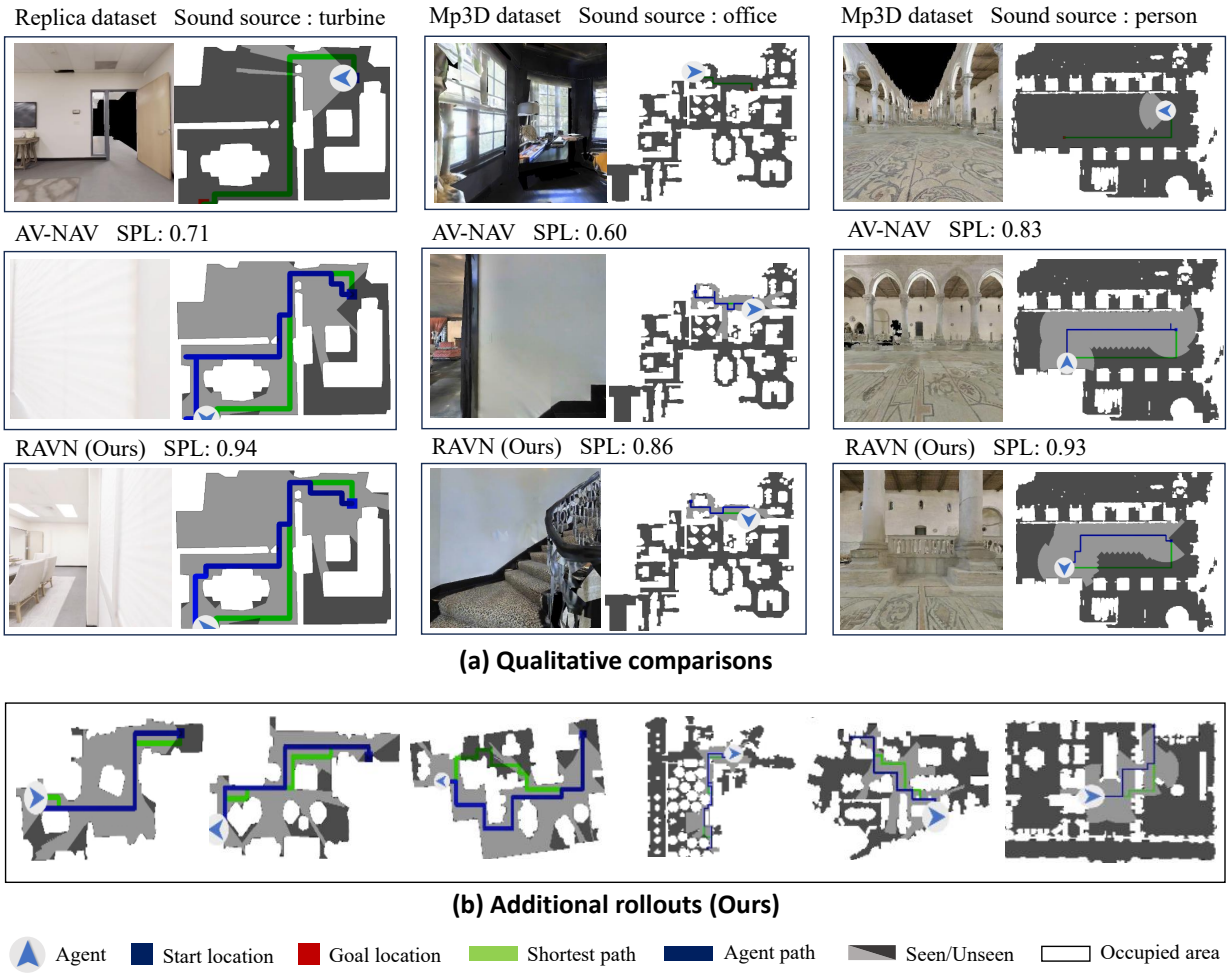

Fig. 4. **Qualitative results.** (a) The top-down path comparisons between the AV-NaV baseline and our RAVN (in some typical Replica and Mp3D rooms). We also put the SPL score here (for each run). (b) Some extra testing paths of RAVN (under different room maps and start-to-goal setups).

D. Qualitative Evaluation

As shown in Fig. 4(a) (from a top-down view), we pick a mix of floor plans (compact, cluttered, and open) from Replica and Matterport3D datasets. RAVN closely follows the shortest path without backtracking or unnecessary detours. This direct behavior is exactly what leads to our SPL gains (as seen earlier). The AV-Nav baseline cannot match this efficiency.

The difference becomes very obvious when the acoustics are difficult. Tricky indoor structures (like sharp corners, long hallways, and blocked doorways) easily cause reverberation and multi-path effects to distort the binaural cues. When the baseline agent relies on these bad signals, it gets lost easily and starts local exploration just to find its direction again. To realize a more stable navigation (than the existing baseline), RAVN passes through these same obstacles without problems. Our reliability-aware fusion keeps the agent stable. Since the model learns to estimate audio reliability through geometric proxies, it simply ignores the misleading sounds that usually confuse regular agents.

Finally, check the extra rollouts in Fig. 4(b). This navigation stability is not a fluke. It persists regardless of the room layout or the specific start-to-goal setup.

E. Ablation Studies

We put all the ablation details in Tables II and III (testing module addition). We start with the AV-Nav baseline first. This baseline just simply combines visual and audio features together (without any geometry or reliability check). To get better results, we add the Acoustic Geometry Reasoner with an MSE loss (+AGR MSE). The navigation performance improves quickly. This directly proves that learning geometry-based representations is very useful.

We also change the MSE loss to a heteroscedastic Negative Log-Likelihood loss (+AGR NLL). The metrics go up even more after this change. We find that modeling the uncertainty (at the observation level) is much better than forcing strict metric accuracy (especially in complex acoustic rooms). Finally, our full RAVN includes the RAGM module. If we drop RAGM from the model (basically going back to the +AGR NLL setup), the performance drops a lot. This big drop means RAGM is very necessary. It uses the learned reliability as an

TABLE II
ABLATION STUDY ON REPLICA DATASET. WE COMPARE DIFFERENT VARIANTS TO VALIDATE THE EFFECTIVENESS OF OUR PROPOSED MODULES.

Model Variant	Replica					
	Heard			Unheard		
	SR(↑)	SPL(↑)	SNA(↑)	SR(↑)	SPL(↑)	SNA(↑)
AV-Nav [1]	93.0	76.1	44.7	47.3	36.8	21.5
+AGR (MSE)	93.7	76.5	49.7	47.7	36.9	21.6
+AGR (NLL)	95.2	79.0	50.1	51.9	39.1	22.1
RAVN (Ours)	**97.0**	**80.2**	**52.4**	**53.1**	**40.3**	**23.4**

TABLE III
ABLATION STUDY ON MP3D DATASET. WE COMPARE DIFFERENT VARIANTS TO VALIDATE THE EFFECTIVENESS OF OUR PROPOSED MODULES.

Model Variant	Matterport3D					
	Heard			Unheard		
	SR(↑)	SPL(↑)	SNA(↑)	SR(↑)	SPL(↑)	SNA(↑)
AV-Nav [1]	68.8	52.3	29.6	33.5	21.9	10.4
+AGR (MSE)	69.4	50.9	27.0	35.3	23.8	12.4
+AGR (NLL)	69.7	54.3	29.4	35.4	24.1	13.0
RAVN (Ours)	**70.9**	**55.6**	**33.2**	**35.6**	**25.9**	**13.6**

adaptive gate to block bad audio signals (ensuring a safe and robust navigation).

V. CONCLUSION

To solve the acoustic ambiguity problem (in spatial navigation), we design the RAVN framework in this paper. We do not treat all the audio signals the same. Instead, our method uses a reliability-aware geometric fusion strategy. We put the Acoustic Geometry Reasoning (AGR) and the RAGM module together to actively drop the misleading sound cues (based on uncertainty estimation). Our tests on Replica and Matterport3D show very good results. RAVN gets much higher success rates and better path efficiency (especially when the agent meets completely unheard sounds).

For future steps, we really want to focus on the sim-to-real transfer. We will put this framework on physical robots (to test its true performance in the real world). We plan to add harsh conditions (like severe background noise and broken sensors) to stress-test the model. Doing this will help us see the actual benefits of our reliability-aware fusion in real life (outside of the simulators).

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] 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.

[3] 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.

[4] 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.

[5] R. Garg, R. Gao, and K. Grauman, "Visually-guided audio spatialization in video with geometry-aware multi-task learning," *International Journal of Computer Vision*, vol. 131, no. 10, pp. 2723–2737, 2023.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[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] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Geometry-aware learning of maps for camera localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2616–2625.

[15] A. Loquercio, M. Segu, and D. Scaramuzza, "A general framework for uncertainty estimation in deep learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3153–3160, 2020.

[16] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu, "Reinforcement learning with unsupervised auxiliary tasks," *arXiv preprint arXiv:1611.05397*, 2016.

[17] 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.

[18] D. Gea and G. Bernardes, "Semantic and spatial sound-object recognition for assistive navigation," in *Proceedings of Conference on Sonification of Health and Environmental Data (SoniHED 2025)*, 2025.

[19] 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.

[20] 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.

[21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[22] 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.

[23] 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.