TRASNAV: TRANSIENT-SOUND-AWARE AUDIO-VISUAL NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-visual navigation has received considerable attention in recent years. However, in realistic multi-source environments, target sounds may occur as brief and temporally sparse acoustic events, providing intermittent and limited evidence for navigation. The key challenge is that these transient target cues can be easily affected by competing sounds, while their short-lived information must also be integrated with navigation context accumulated over much longer time scales. To address this problem, we propose TraSNav, a navigation framework for transient target sounds. Our key idea is to maintain an event-level memory that accumulates and consolidates information from fragmented acoustic observations to improve robustness against competing sounds, and then adaptively fuse this event memory with long-term navigation memory to handle temporal-scale differences. Experimental results demonstrate that TraSNav consistently outperforms representative audio-visual navigation methods, achieving up to a 26.41% relative improvement in success rate when each target-sound event is shortened to 10% of its original duration in multi-source environments.

1 INTRODUCTION

Embodied navigation enables autonomous agents to perceive unfamiliar environments and reach designated targets through sequential interaction, and has been widely investigated in PointGoal navigation (Savva et al., 2019), ObjectGoal navigation (Zhang et al., 2021; Majumdar et al., 2022), and vision-and-language navigation (Anderson et al., 2018b). However, visual perception is inherently constrained by occlusion and limited field of view, making targets outside the current view difficult to discover efficiently. To complement visual perception, audio-visual navigation equips embodied agents with auditory capabilities and requires them to locate sounding targets by jointly exploiting visual observations and binaural acoustic cues (Gan et al., 2020; Chen et al., 2020; 2021b;a).

As audio-visual navigation has progressed, many studies have shifted from simple single-source settings to more realistic multi-source scenarios (Yu et al., 2022; Younes et al., 2023; Chen et al., 2023; Shi et al., 2025). Meanwhile, some works have considered intermittent target sounds, mainly focusing on when target sounds are available and how previously observed acoustic information can be utilized during silent periods (Chen et al., 2021a; Liu et al., 2024; Zeng et al., 2026). The reliability of each individual acoustic observation, however, has received less attention. This issue becomes particularly important in realistic multi-source environments, where brief and sparse target sounds can be easily affected by competing sources. As shown in Fig. 1, when each target sound is shortened to 10% of its original duration under multi-source settings, the performance of state-of-the-art methods drops substantially.

We refer to such brief and sparse target sounds as acoustic events. Effectively using these acoustic events in multi-source environments introduces two main challenges. First, each short acoustic event typically provides only partial target evidence. Competing sources further increase the ambiguity of individual observations, causing target cues collected at different times to vary in reliability or even conflict with one another. Therefore, a single event is often insufficient to form a stable and reliable target representation, requiring evidence to be accumulated across multiple events. Second, event-level acoustic evidence and navigation context evolve at substantially different temporal scales. Target-related acoustic information is available only during a limited number of brief events, whereas visual, spatial, and trajectory information continuously accumulates throughout navigation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

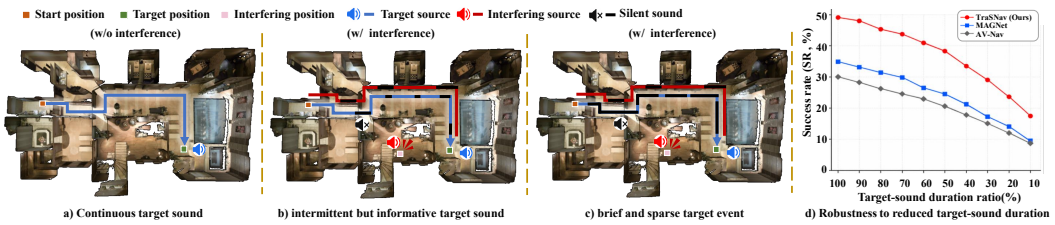


Figure 1: Target-sound conditions in audio-visual navigation. AV-Nav (Chen et al., 2020) and MAGNet (Zeng et al., 2026) represent (a) continuously available and (b) intermittently available target sounds. They mainly focus on temporal availability, while the reliability of brief target-sound events under competing-source conditions is less explored, leading to substantial performance degradation as target-sound events are progressively shortened.

Effectively integrating such short-lived target evidence with increasingly rich long-term navigation context therefore remains challenging.

To address these challenges, we propose **TraSNav, a transient-sound-aware audio-visual navigation framework for multi-source environments.** TraSNav consists of two key components: RTEM first generates multiple candidate target hypotheses from the current acoustic observation, and then verifies and refines them using recent supporting evidence and historical conflicting evidence. In this way, RTEM builds a more reliable target-event representation from distributed event-level observations. Based on this representation, TSAMF adaptively controls how much the verified event information contributes to navigation memory according to the current event state and accumulated temporal context. This allows transient acoustic cues to be effectively integrated with continuously accumulated long-term scene information for long-horizon navigation.

Our main contributions are summarized as follows:

- We establish a controlled transient-sound evaluation protocol by progressively shortening individual target-sound events while preserving their original occurrence patterns, source configurations, and navigation conditions. This protocol enables systematic evaluation under increasingly limited event-level acoustic evidence.
- We propose TraSNav, a transient-sound-aware audio-visual navigation framework for multi-source environments. It builds a robust event-level memory for brief and sparse acoustic cues and adaptively fuses it with long-term navigation memory to handle competing sounds and temporal-scale differences.
- Extensive experiments under both original and progressively shortened target-sound conditions demonstrate that TraSNav consistently improves navigation success and efficiency, with stronger robustness as the acoustic evidence available in each target event becomes increasingly limited.

2 Related Work

2.1 Embodied Navigation

Embodied navigation requires an autonomous agent to perceive unseen environments and reach designated goals through sequential interactions. Existing studies have explored a variety of navigation tasks, including PointGoal navigation (Anderson et al., 2018a; Partsey et al., 2022), ObjectGoal navigation (Zhang et al., 2021; Ramakrishnan et al., 2022; Majumdar et al., 2022), and vision-and-language navigation (Anderson et al., 2018b; Chen et al., 2021c; Li & Bansal, 2023; Gao et al., 2023). These studies have advanced embodied navigation across geometric goals, semantic targets, and language-guided tasks. However, these paradigms predominantly rely on visual observations, which are inherently constrained by occlusion and limited field of view. In contrast, auditory sensing can provide directional and spatial cues for targets beyond the current visual range, motivating its increasing use in embodied navigation and the development of audio-visual navigation.

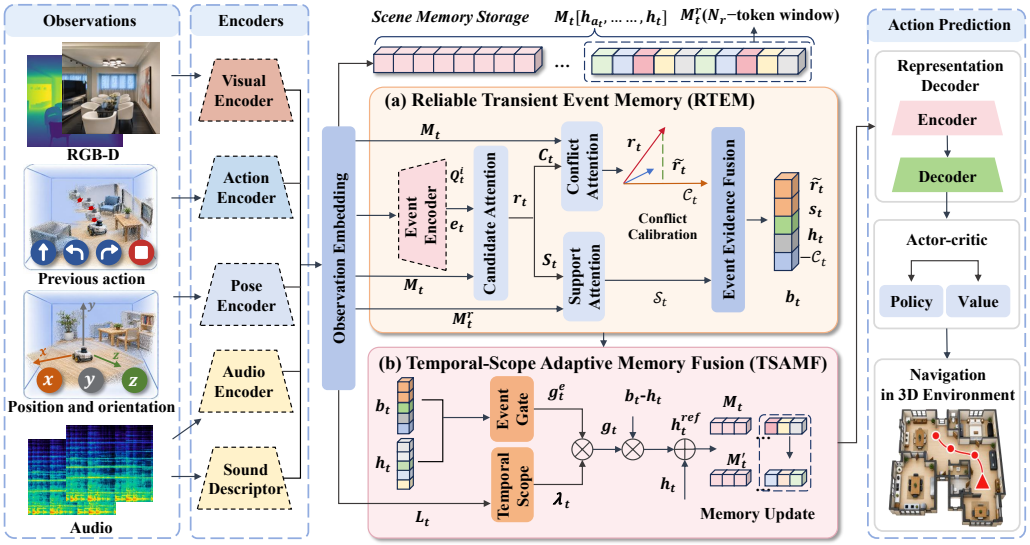


Figure 2: Overall framework of TraSNav. The multi-modal observations are encoded and maintained in the temporal scene memory. Reliable Transient Event Memory (RTEM) constructs a verified target-event representation by reasoning over candidate hypotheses, recent supporting evidence, and historical conflicting evidence. Temporal-Scope Adaptive Memory Fusion (TSAMF) further adjusts the contribution of the verified event according to the current navigation context and memory length, and updates the current memory state for subsequent navigation decision.

2.2 Audio-Visual Navigation

Early audio-visual navigation approaches mainly considered continuously available target sounds. AV-Nav learns an end-to-end policy for navigation action prediction (Chen et al., 2020), while AV-WaN further introduces navigable waypoint prediction and spatially grounded acoustic memory (Chen et al., 2021b), establishing the basic audio-visual navigation paradigm.

As research moves toward more realistic acoustic environments, subsequent studies have considered intermittent sounds, environmental noise, dynamic sources, and more flexible information acquisition. SAVi maintains target information under sporadic sounds through persistent multimodal memory and goal descriptors (Chen et al., 2021a). SAAVN (Yu et al., 2022) and DAV-Nav (Younes et al., 2023) further investigate navigation under noisy competing sources and moving sound sources, respectively. ORAN (Chen et al., 2023) and CAVEN (Liu et al., 2024) explore enhanced perception and interactive assistance for more flexible audio-visual navigation. More recently, audio-visual navigation has been extended to multi-source and continuous environments. ENMuS^3 models complex acoustic scenes with sound event descriptors and multi-scale scene memory transformers (Shi et al., 2025), while MAGNet leverages historical context and self-motion information to maintain goal awareness when target sounds become intermittently unavailable in continuous environments (Zeng et al., 2026).

Despite the advances, the problem of modeling brief and sparse acoustic events in multi-source environments remains less explored, although such cases are common in realistic navigation scenarios.

3 Method

3.1 Framework Overview

As illustrated in Fig. 2, TraSNav is designed to handle brief and low-information target sounds in multi-source environments by introducing an event-level memory and integrating it with temporal-level navigation memory. For the temporal-level memory, we follow ENMuS^3 (Shi et al., 2025) and

fuse visual, acoustic, target-category, GoalDescriptor, action, and pose information at each time step to obtain the current observation representation h_t. The temporal scene memory is then defined as $\mathcal{M}_t = [h_{a_t}, \ldots, h_t]$, where $a_t = \max(1, t - C)$ and C denotes the predefined memory length.

For the event-level memory, we introduce **Reliable Transient Event Memory (RTEM)**, which reasons over multiple target hypotheses and evaluates them using recent supporting context and historical conflict-sensitive evidence to obtain a reliable target-event representation. Based on this representation, **Temporal-Scope Adaptive Memory Fusion (TSAMF)** adaptively controls its contribution to the temporal memory according to the current event information and accumulated temporal context. The resulting fused memory is finally processed by the multi-scale scene memory transformer (MSMT) for action prediction. We next describe these components in detail.

3.2 Reliable Transient Event Memory

RTEM aims to derive reliable target-event information from multiple brief and intermittent acoustic observations. To this end, it introduces three sets of learnable queries: target-hypothesis queries $\{q_i\}_{i=1}^{N_Q}$, supporting-cue queries $\{s_i\}_{i=1}^{N_S}$, and conflict-evidence queries $\{c_i\}_{i=1}^{N_C}$. These queries retrieve complementary event-level evidence from the current observation h_t and the temporal scene memory $\mathcal{M}_t$, enabling RTEM to construct a more reliable representation of transient target sounds. In practice, we set $N_Q = 3$, $N_S = 1$, and $N_C = 1$. We omit the subscript of s_i and c_i for simplicity.

Initial event representation. The target-hypothesis queries are used to retrieve target-related event evidence from the temporal scene memory. We first construct the current event feature as

$$\begin{aligned} e_t &= \phi_e(h_t), \\ Q_t^i &= q_i + \phi_q\left([e_t, h_t]\right), \\ u_t^i &= \mathrm{TFBlock}\left(Q = Q_t^i,\, K = \mathcal{M}_t,\, V = \mathcal{M}_t\right), \qquad i = 1, \ldots, N_Q. \end{aligned} \tag{1}$$

where $\phi_\cdot(\cdot)$ denotes an MLP-style block, with the detailed configurations of all such blocks provided in *Suppl.* Table S2, $[\cdot]$ denotes feature concatenation along the channel dimension, and $\mathrm{TFblock}(\cdot)$ denotes the Transformer-based retrieval block. In this way, each u_t^i encodes a target hypothesis retrieved from the temporal scene memory. We further aggregate these hypotheses as

$$\alpha_t^i = \frac{\exp\left(\phi_\alpha\left([u_t^i, e_t, h_t]\right)\right)}{\sum_{j=1}^{N_Q} \exp\left(\phi_\alpha\left([u_t^j, e_t, h_t]\right)\right)}, \quad r_t^{\mathrm{raw}} = \sum_{i=1}^{N_Q} \alpha_t^i u_t^i, \quad r_t = \mathrm{LN}\left(r_t^{\mathrm{raw}} + \phi_r\left([r_t^{\mathrm{raw}}, e_t, h_t]\right)\right), \tag{2}$$

where $\odot$ denotes element-wise multiplication, and $\mathrm{LN}(\cdot)$ denotes layer normalization. We denote r_t as the initial sound-event representation and further evaluate its reliability using supportive and conflicting evidence retrieved by the other two query sets.

Supportive evidence retrieval. We retrieve the supportive cue from a local temporal window:

$$S_t = s + \phi_s\left([r_t, e_t, h_t]\right), \qquad \mathcal{S}_t = \mathrm{TFBlock}\left(Q = S_t,\, K = \mathcal{M}_t^r,\, V = \mathcal{M}_t^r\right), \tag{3}$$

where $\mathcal{M}_t^r = \left[h_{\max(a_t, t-N_r+1)}, \ldots, h_t\right]$ denotes the recent temporal memory with window length N_r. Unavailable padded positions are masked out during attention.

Conflicting evidence retrieval. We retrieve conflicting evidence from the full temporal memory:

$$C_t = c + \phi_c\left([r_t, e_t, h_t, |r_t - e_t|]\right), \quad \mathcal{C}_t = \mathrm{TFBlock}\left(Q = C_t,\, K = \mathcal{M}_t,\, V = \mathcal{M}_t\right), \tag{4}$$

Here, we explicitly include $|r_t - e_t|$ as a discrepancy cue to provide an inductive bias for identifying information that is inconsistent with the current event representation. $\mathcal{C}_t$ then captures conflicting evidence between the current event representation r_t and historical observations. We therefore use it to calibrate r_t by

$$\eta_t = \mathrm{sigmoid}\left(\phi_\eta\left([r_t, e_t, \mathcal{C}_t]\right)\right), \quad \tilde{r}_t = \mathrm{LN}\left(r_t - \eta_t \frac{r_t^\top \mathcal{C}_t}{\max(\|\mathcal{C}_t\|_2^2, \epsilon_{\mathrm{proj}})} \mathcal{C}_t\right), \tag{5}$$

where ϵ_{proj} is a small constant for numerical stability. The gate η_t controls the calibration strength, while the projection term suppresses the component of r_t aligned with the conflicting evidence.

Context-dependent event refinement. Finally, a context-dependent weighting function estimates the relative contributions of the available evidence:

$$\omega_t = \mathrm{softmax}\left(\phi_\omega([\tilde{r}_t, h_t, \mathcal{S}_t, \mathcal{C}_t])\right), \quad b_t^{\mathrm{raw}} = \omega_t^1 \tilde{r}_t + \omega_t^2 \mathcal{S}_t + \omega_t^3 h_t - \omega_t^4 \mathcal{C}_t, \tag{6}$$

where $\omega_t \in \mathbb{R}^4$, $\omega_t^k \geq 0$, and $\sum_{k=1}^4 \omega_t^k = 1$. The four weights correspond to the calibrated target representation $\tilde{r}_t$, supportive temporal evidence s_t, current observation h_t, and conflicting evidence $\mathcal{C}_t$, respectively. The final event representation is then obtained as

$$b_t = \mathrm{LN}(\mathrm{LN}(b_t^{\mathrm{raw}}) + \mathrm{FFN}(\mathrm{LN}(b_t^{\mathrm{raw}}))), \tag{7}$$

where $\mathrm{FFN}(\cdot)$ denotes a feed-forward network. The resulting b_t serves as the verified target-event representation for subsequent temporal fusion.

3.3 Temporal-Scope Adaptive Memory Fusion

We next fuse the temporal-level memory with the event-level memory. To this end, two factors should be considered. First, the fusion result should be jointly determined by the current observation h_t and the verified event representation b_t. We therefore introduce an event-conditioned gate:

$$g_t^e = g_{\max} \cdot \mathrm{sigmoid}\left(\phi_g\left([h_t, b_t, |b_t - h_t|]\right)\right) \in \mathbb{R}, \tag{8}$$

where $g_{\max}$ denotes the maximum value of the event-conditioned gate.

Second, the contribution of the event representation should also depend on the amount of temporal context accumulated during navigation. At early navigation stages, the temporal memory contains limited information, and event-level acoustic evidence can provide important guidance. As navigation proceeds, increasingly rich visual, spatial, and trajectory information is accumulated in the temporal memory. In this case, the influence of transient event information should gradually decrease to avoid overwriting the more stable long-term navigation context. We therefore introduce a temporal-scope factor:

$$\lambda_t = \lambda_{\min} + (1 - \lambda_{\min}) \cdot \mathrm{sigmoid}\left(\frac{L_s - L_t}{\tau}\right), \qquad g_t = g_t^e \lambda_t, \tag{9}$$

where L_t denotes the amount of temporal context accumulated at time step t, L_s controls the transition point, and τ controls the smoothness of the decay. As L_t increases, λ_t gradually decreases toward $\lambda_{\min}$, reducing the influence of transient event information as the long-term navigation context becomes richer. Detailed visualization of the two-stage fusion process is provided in ***Suppl.*** Fig. S3. The resulting fused scene representation and event-enhanced temporal memory $\mathcal{M}_t'$, which are subsequently used for action prediction, are obtained as

$$h_t^{\mathrm{ref}} = h_t + g_t(b_t - h_t), \qquad \mathcal{M}_t' = [h_{a_t}, \ldots, h_{t-1}, h_t^{\mathrm{ref}}]. \tag{10}$$

4 Experiments

4.1 Experimental Setup

Environments and simulator. We use the ENMuS^3 multi-source audio-visual navigation dataset (Shi et al., 2025) for training and evaluation. ENMuS^3 is built on indoor scenes from Matterport3D (Chang et al., 2017) and simulates audio-visual observations using SoundSpaces (Chen et al., 2020), which provides egocentric visual observations and spatialized audio based on geometry-aware room impulse responses. In addition to the target sound, interfering sounds are introduced to construct multi-source acoustic environments. The dataset contains 440,238, 10,000, and 9,873 episodes for training, validation, and testing, respectively, with disjoint Matterport3D scenes across the three subsets.

Scenario configurations. We analyze the sound activity duration in the ENMuS^3 test split. Under the original setting, the target sound is active for approximately 44.61% of an episode on average, while interfering sounds are active for approximately 44.11%. We conduct experiments under the original target-sound setting and nine transient settings, where the active duration of each target-sound event is shortened to 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, and 10% of its original

duration, while the interfering sounds remain unchanged. Consequently, under the most challenging 10% setting, the target sound is active for only approximately 4.46% of the episode on average, providing highly limited acoustic evidence for navigation.

Evaluation metrics. Similar to existing audio-visual navigation studies (Chen et al., 2020; 2021b; Shi et al., 2025), the following metrics are adopted to evaluate the performance of different navigation methods: (1) success rate (SR), (2) success weighted by inverse path length (SPL) (Anderson et al., 2018a), (3) success weighted by inverse number of actions (SNA) (Chen et al., 2021b), and (4) average distance to goal when an episode terminates (DTG). Higher SR, SPL, and SNA indicate better navigation performance, whereas a lower DTG is preferred.

Baseline methods. We conduct a comparative analysis of our proposed TraSNav against the following representative audio-visual navigation methods.

- **AV-Nav** (Chen et al., 2020): An end-to-end reinforcement learning policy that encodes audio and visual observations with a GRU and directly predicts low-level navigation actions.
- **SAVi** (Chen et al., 2021a): An end-to-end audio-visual navigation policy that adopts a ResNet to estimate the location of the target sound source and makes navigation decisions using a transformer-based architecture.
- **SMT + Audio** (Fang et al., 2019): A transformer-based navigation policy that encodes historical scene observations into a memory representation and predicts the next navigation action by decoding this representation. Following previous audio-visual navigation settings, audio observations are additionally incorporated into its memory representation.
- **ENMuS^3** (Shi et al., 2025): A multi-source audio-visual navigation framework that employs a sound event descriptor to extract target-related acoustic information and a scene memory transformer to aggregate multimodal observations over time.
- **MAGNet** (Zeng et al., 2026): A memory-augmented audio-visual navigation model that jointly represents spatial and semantic target information and incorporates historical context and self-motion cues for goal reasoning and navigation.

For a fair comparison, all methods use the same train/validation/test split. For each transient-sound condition, the same target-sound shortening setting is applied to all methods. Across different acoustic conditions, the number and onset times of target-sound events remain unchanged, while their active durations are progressively shortened, resulting in longer silent intervals. The scene, initial agent pose, navigation target, and other episode configurations are kept unchanged. Where applicable, all baseline methods and our TraSNav also use the same reward function and input configurations, including RGB-D observations with a resolution of 128×128 pixels and audio waveforms sampled at 16,000 Hz. More implementation details can be found in *Suppl.* Table S1.

4.2 Comparative Experiments

Quantitative experimental results. Table 1 presents the quantitative comparison of TraSNav with representative audio-visual navigation methods under the original target-sound setting. The best results are highlighted in bold. Among the methods with available results, TraSNav achieves the best performance across all four evaluation metrics. Specifically, TraSNav achieves an SR of 48.94%, outperforming the strongest baseline ENMuS^3 by 2.98 percentage points. Meanwhile, TraSNav improves SPL and SNA by 2.35 and 2.80 percentage points, respectively, while reducing DTG from 6.1867 to 5.9532. These results demonstrate that TraSNav improves both navigation success and navigation efficiency in multi-source environments.

To further evaluate robustness to transient target sounds, Table 2 reports the results as the target-sound duration is progressively reduced from 90% to 10% of its original duration. Among the methods with available results, TraSNav consistently achieves the best SR and SPL across all the transient-sound conditions. Compared with ENMuS^3, TraSNav improves SR by 3.46, 5.06, and 4.87 percentage points under the 80%, 50%, and 20% settings, respectively. As the target-sound duration decreases, the navigation performance of existing methods degrades to varying degrees, whereas TraSNav consistently maintains superior performance. These results indicate that TraSNav exhibits stronger robustness when the available target acoustic evidence becomes increasingly brief and temporally sparse.

Table 1: Quantitative comparison with representative audio-visual navigation methods under the original target-sound setting. Best in bold.

Method	SR↑	SPL↑	SNA↑	DTG↓
AV-Nav (Chen et al., 2020)	30.04	16.51	20.95	8.72
SAVi (Chen et al., 2021a)	30.26	15.67	19.87	8.38
SMT + Audio (Fang et al., 2019)	38.21	19.52	27.19	6.77
ENMuS^3 (Shi et al., 2025)	45.96	24.49	34.03	6.19
MAGNet (Zeng et al., 2026)	34.82	20.87	28.54	7.57
TraSNav (Ours)	**48.94**	**26.84**	**36.83**	**5.95**

Table 2: Quantitative comparison across different transient target-sound duration ratios. SR and SPL are reported for all methods. Best in bold.

Sound Ratio	AV-Nav (Chen et al., 2020)		SAVi (Chen et al., 2021a)		SMT + Audio (Fang et al., 2019)		ENMuS^3 (Shi et al., 2025)		MAGNet (Zeng et al., 2026)		TraSNav (Ours)	
	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
90%	28.24	15.23	28.71	14.95	36.25	17.08	44.28	23.51	32.99	19.91	**48.05**	**26.05**
80%	26.19	14.05	27.18	13.78	33.88	15.77	41.77	22.35	31.29	18.55	**45.23**	**24.32**
70%	24.58	13.06	25.89	12.72	30.26	14.27	39.20	20.80	29.80	17.53	**43.83**	**23.13**
60%	22.95	11.89	22.59	10.79	28.23	13.98	36.48	18.90	26.46	15.71	**40.96**	**21.26**
50%	20.71	10.66	20.54	9.80	25.62	12.23	33.30	16.64	24.46	14.23	**38.36**	**19.62**
40%	17.94	9.32	18.06	8.37	22.34	10.87	28.73	14.25	21.23	12.11	**33.42**	**16.40**
30%	15.04	7.60	14.83	6.59	18.55	8.61	24.37	11.56	17.29	9.82	**28.98**	**13.94**
20%	12.02	5.84	11.20	5.10	15.65	6.12	18.67	8.62	13.95	7.97	**23.54**	**10.66**
10%	8.87	4.34	8.55	3.73	11.12	5.23	13.82	6.35	9.26	5.39	**17.47**	**7.32**

Qualitative experimental results. Fig. 3 shows the navigation trajectories of TraSNav and representative audio-visual navigation methods on the same test episode under representative target-sound durations. Under the original sound condition, the compared methods are able to successfully complete the navigation task with stable trajectories. But as the target-sound duration is progressively reduced, the navigation trajectories of existing methods become increasingly unstable, and several methods finally fail to reach the target under more challenging transient-sound conditions. In contrast, TraSNav successfully reaches the target under all four acoustic conditions and maintains relatively stable navigation behavior even when the available target acoustic evidence becomes substantially reduced. This qualitative observation is consistent with the quantitative results, illustrating the ability of TraSNav to exploit brief and temporally sparse acoustic evidence for reliable navigation in multi-source environments.

4.3 Ablation Study

To analyze the contributions of different designs in TraSNav to transient-sound navigation, we conduct ablations from two aspects: **event evidence verification** and **temporal-scope adaptive fusion**. We first examine the roles of different evidence branches within RTEM. **w/o Support** removes the retrieval and utilization of recent supporting evidence, while retaining candidate reasoning, historical conflict evidence, and the subsequent TSAMF. **w/o Conflict** removes historical conflict retrieval and conflict calibration, while preserving candidate reasoning, recent supporting evidence, and TSAMF. These two variants allow us to separately investigate the contributions of recent consistent evidence and historical conflicting information to reliable target-event modeling.

We further introduce **RTEM-only**, which retains the complete RTEM, including multi-candidate target reasoning, recent supporting evidence, historical conflicting evidence, and conflict calibration, while disabling TSAMF. In this variant, the verified event information is no longer adaptively integrated with long-term scene memory according to the temporal scope. This setting therefore separates the contribution of **reliable event evidence modeling** from that of **adaptive memory fusion**. The full TraSNav enables both the complete evidence reasoning process in RTEM and TSAMF, allowing us to evaluate their complementary effects on overall navigation performance.

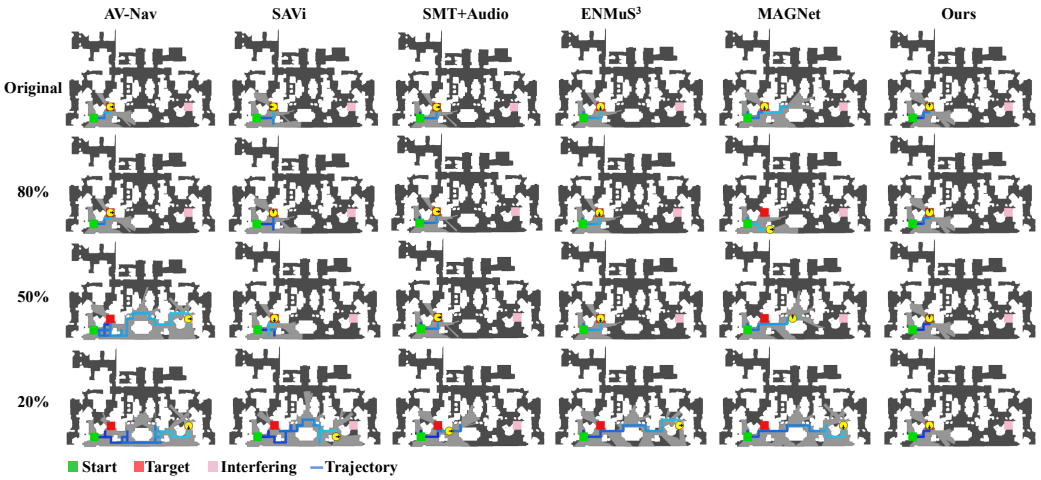


Figure 3: Qualitative comparison under representative progressively shortened target-sound conditions. Navigation trajectories of different methods are visualized on the same test episode under the Original, 80%, 50%, and 20% target-sound conditions. As the target-sound duration decreases, existing methods exhibit increasingly unstable navigation behavior or fail to reach the target, whereas TraSNav successfully completes the navigation task across all four conditions.

Table 3: Component ablation under the original target-sound setting. Best in bold.

Variant	Components				Metrics			
	Cand.	Supp.	Conf.	TSAMF	SR↑	SPL↑	SNA↑	DTG↓
Baseline	×	×	×	×	45.96	24.49	34.03	6.19
w/o Support	✓	×	✓	✓	47.05	25.32	36.21	6.01
w/o Conflict	✓	✓	×	✓	47.23	26.11	36.72	**5.95**
RTEM-only	✓	✓	✓	×	46.28	24.86	34.59	6.13
TraSNav	✓	✓	✓	✓	**48.94**	**26.84**	**36.83**	**5.95**

Table 3 summarizes the component ablations under the original target-sound setting. Removing either supporting or conflicting evidence degrades the overall performance, confirming that both evidence branches contribute to reliable target-event verification. In particular, removing recent support results in a larger degradation than removing conflict evidence on SR, SPL, and SNA, indicating the importance of recent consistent observations for stabilizing target estimation. Moreover, compared with **RTEM-only**, the full TraSNav improves SR from 46.28 to 48.94 and SPL from 24.86 to 26.84, showing that temporal-scope adaptive fusion provides additional benefits beyond reliable event modeling alone.

Table 4 evaluates the ablation variants under progressively shortened target sounds. TraSNav consistently achieves the best SR and SPL across all three conditions. Notably, its SR improvements over **RTEM-only** increase from 2.65 percentage points at 80% to 3.14 and 3.56 percentage points at 50% and 20%, respectively. The widening performance gap suggests that temporal-scope adaptive fusion becomes increasingly beneficial as the available acoustic evidence becomes more limited. The consistent degradation of **w/o Support** and **w/o Conflict** further indicates that the two evidence branches remain complementary under transient-sound conditions.

Table 5 further isolates candidate reasoning, recent support, and conflict reasoning under different target-sound conditions. Both evidence branches consistently improve over candidate reasoning alone, while their combination achieves the best overall performance, confirming that recent supportive and historical conflicting evidence provide complementary information for reliable event estimation.

Table 4: Ablation results under different transient target-sound conditions. Best in bold.

Variant	80%		50%		20%	
	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
Baseline	41.77	22.35	33.30	16.64	18.67	8.62
w/o Support	43.85	23.33	36.59	17.98	20.37	9.02
w/o Conflict	44.34	24.10	37.56	19.22	21.56	9.77
RTEM-only	42.58	22.78	35.22	17.68	19.98	9.48
TraSNav	**45.23**	**24.32**	**38.36**	**19.62**	**23.54**	**10.66**

Table 5: Component analysis of RTEM under different target-sound conditions. TSAMF is disabled for all variants to isolate the contribution of RTEM. Best in bold.

Variant	RTEM Components			Original		80%		50%		20%	
	Cand.	Supp.	Conf.	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
Baseline	×	×	×	45.96	24.49	41.77	22.35	33.30	16.64	18.67	8.62
Candidate only	✓	×	×	46.01	24.51	41.94	22.46	33.52	16.90	19.02	9.45
Cand. + Support	✓	✓	×	46.18	24.66	**42.59**	22.60	34.97	17.23	19.52	9.21
Cand. + Conflict	✓	×	✓	46.13	24.63	42.21	22.53	34.23	17.11	19.44	9.03
Full RTEM	✓	✓	✓	**46.28**	**24.86**	42.58	**22.78**	**35.22**	**17.68**	**19.98**	**9.48**

Table 6: Sensitivity to the number of RTEM target-hypothesis queries N_Q under different target-sound conditions.

N_Q	Original		80%		50%		20%	
	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
1	48.21	26.27	44.43	23.65	36.88	18.23	21.27	10.02
2	48.57	26.46	44.77	24.01	37.72	19.22	22.86	10.10
3	**48.94**	**26.84**	**45.23**	**24.32**	**38.36**	**19.62**	**23.54**	**10.66**
4	48.61	26.53	44.85	24.22	37.96	18.97	23.12	10.50

We further examine the effect of the number of target-hypothesis queries N_Q in Table 6. Increasing N_Q from 1 to 3 consistently improves both SR and SPL across the original and transient target-sound conditions. The improvement is particularly evident under the challenging 20% condition, where SR increases from 21.27% to 23.54%. Increasing N_Q to 4 provides no further benefit, suggesting that additional hypotheses beyond three mainly introduce redundant information. We therefore use $N_Q = 3$ throughout all experiments.

5 Conclusion

In this work, we study audio-visual navigation with brief and intermittent target sounds in multi-source environments. To address unreliable target-event estimation and the temporal mismatch between transient acoustic evidence and long-term navigation context, we propose TraSNav, a transient-sound-aware navigation framework. TraSNav introduces Reliable Transient Event Memory (RTEM) to construct robust target-event representations by consolidating complementary event-level evidence, and Temporal-Scope Adaptive Memory Fusion (TSAMF) to adaptively integrate verified event information with accumulated navigation context. We further establish controlled transient-sound settings by progressively shortening target-sound events while preserving their occurrence patterns and interference conditions. Extensive experiments demonstrate that TraSNav consistently improves navigation performance under both original and shortened target-sound conditions, with larger gains as the available target acoustic evidence becomes more limited. Ablation studies further verify the complementary contributions of RTEM and TSAMF. Future work will extend TraSNav from simulated indoor environments to real-world scenarios with more diverse acoustic dynamics and sensing uncertainties.

AI USE STATEMENT

During the preparation of this manuscript, the authors used generative AI tools to assist with language refinement, formatting checks, and optimization of technical descriptions. All research contributions, methodology designs, experimental procedures, result analyses, and scientific conclusions were developed by the authors and independently verified. The authors take full responsibility for the accuracy and completeness of this manuscript.

REFERENCES

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018a.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683, 2018b.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *IEEE International Conference on 3D Vision (3DV)*, pp. 667–676, 2017.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-visual navigation in 3D environments. In *European Conference on Computer Vision (ECCV)*, pp. 17–36, 2020.

Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15516–15525, 2021a.

Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *International Conference on Learning Representations*, 2021b.

Jinyu Chen, Wenguan Wang, Si Liu, Hongsheng Li, and Yi Yang. Omnidirectional information gathering for knowledge transfer-based audio-visual navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10993–11003, 2023.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5834–5847, 2021c.

Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 538–547, 2019.

Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B. Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9701–9707, 2020.

Chen Gao, Xingyu Peng, Mi Yan, He Wang, Lirong Yang, Haibing Ren, Hongsheng Li, and Si Liu. Adaptive zone-aware hierarchical planner for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14911–14920, 2023.

Jialu Li and Mohit Bansal. Improving vision-and-language navigation by generating future-view image semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10803–10812, 2023.

Xiulong Liu, Sudipta Paul, Moitreya Chatterjee, and Anoop Cherian. CAVEN: An embodied conversational agent for efficient audio-visual navigation in noisy environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3765–3773, 2024.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. ZSON: Zero-shot object-goal navigation using multimodal goal embeddings. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32340–32352, 2022.

Ruslan Partsey, Erik Wijmans, Naoki Yokoyama, Oles Dobosevych, Dhruv Batra, and Oleksandr Maksymets. Is mapping necessary for realistic pointgoal navigation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17232–17241, 2022.

Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. PONI: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18890–18900, 2022.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9339–9347, 2019.

Zhanbo Shi, Lin Zhang, Linfei Li, and Ying Shen. Towards audio-visual navigation in noisy environments: A large-scale benchmark dataset and an architecture considering multiple sound-sources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14673–14680, 2025.

Abdelrahman Younes, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds. *IEEE Robotics and Automation Letters*, 8(2):928–935, 2023.

Yinfeng Yu, Wenbing Huang, Fuchun Sun, Changan Chen, Yikai Wang, and Xiaohong Liu. Sound adversarial audio-visual navigation. In *International Conference on Learning Representations*, 2022.

Yichen Zeng, Hebaixu Wang, Meng Liu, Yu Zhou, Chen Gao, Kehan Chen, and Gongping Huang. Semantic audio-visual navigation in continuous environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22369–22379, 2026.

Sixian Zhang, Xinhang Song, Yubing Bai, Weijie Li, Yakui Chu, and Shuqiang Jiang. Hierarchical object-to-zone graph for object navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15130–15140, 2021.

A Appendix

This appendix provides additional implementation details, extended experiments, and qualitative analyses to complement the method description and experimental results in the main paper, including:

- **Appendix A.1:** implementation and training details of TraSNav, including RTEM, TSAMF, and the overall navigation policy.
- **Appendix A.2:** evaluation protocol and baseline adaptation, including transient-sound construction and the unified evaluation setting.
- **Appendix A.3:** mechanism analysis of RTEM and TSAMF, including historical evidence retrieval and adaptive temporal fusion.
- **Appendix A.4:** training stability analysis of TraSNav across multiple independent runs.
- **Appendix A.5:** additional episode-level trajectory comparisons under different target-sound durations.
- **Appendix A.6:** limitations and potential directions for future work.

Table S1: Implementation and training configurations of TraSNav.

Module	Configuration	Setting
Navigation	Multimodal observation	RGB-D (128), audio (128), target category (20), GoalDescriptor (128), previous action (16), and relative pose (4)
	Observation dimension	424
	Hidden / memory representation	$D = 256$; scene-memory backbone: MSMT
RTEM	Query configuration	$N_Q = 3$ target-hypothesis queries; $N_S = 1$ support query; $N_C = 1$ conflict query
	Transformer retrieval	8 attention heads; FFN expansion ratio $= 4$
	Recent-memory window	$N_r = 28$
	Historical memory capacity	159 historical tokens (160 positions including the current token)
	Numerical stability	Projection denominator floor $\epsilon_{\text{proj}} = 10^{-6}$
	Regularization	Dropout $= 0$
TSAMF	Event gate	$g_{\max} = 0.35$
	Temporal-scope modulation	$\lambda_{\min} = 0.12$; $L_s = 80$; $\tau = 25$
	Refinement scope	Current observation token only
Optimization	Optimizer	Adam; learning rate $= 2.5 \times 10^{-4}$; $\epsilon_{\text{Adam}} = 10^{-5}$
	PPO configuration	Clipping parameter $= 0.2$; PPO epochs $= 2$; mini-batches $= 2$
	Rollout	160 steps per environment; 32 parallel environments
	Loss coefficients	Value loss $= 0.5$; entropy $= 0.05$; maximum gradient norm $= 0.2$
	Advantage estimation	$\gamma = 0.99$; GAE coefficient $= 0.95$
	Scheduling	Linear learning-rate and clipping decay disabled
	Training budget	60,000 policy updates; approximately 3.072×10^8 environment interactions
	Checkpointing	Saved every 50 updates; selected by highest validation SPL

A.1 Implementation Details

TraSNav adopts a temporal scene-memory architecture for audio-visual navigation. At each time step, RGB-D observations, current acoustic observations, target category, GoalDescriptor, the previous action, and relative pose form a 424-dimensional multimodal observation, which is encoded into a $D = 256$ dimensional current representation h_t. Historical observations form the temporal scene memory modeled by MSMT. Based on this memory, RTEM uses three query sets for target-hypothesis retrieval, recent supportive-evidence retrieval, and full-history conflict-evidence retrieval, and further constructs a conflict-calibrated event representation. TSAMF then adaptively regulates the contribution of this event representation according to the current event reliability and the amount of temporal context accumulated during navigation. The overall implementation and training configurations are summarized in Table S1, and the architectures of the MLP-style mappings $\phi_\cdot$ are summarized in Table S2.

All internal representations in RTEM and TSAMF use a feature dimension of $D = 256$. The three Transformer-based retrieval branches differ in their query semantics and temporal scopes. For target-hypothesis retrieval, $Q = Q_t^i$ and $K = V = \mathcal{M}_t$; for supportive-evidence retrieval, $Q = S_t$ and $K = V = \mathcal{M}_t^r$; and for conflict-evidence retrieval, $Q = C_t$ and $K = V = \mathcal{M}_t$. All three retrieval blocks employ eight-head multi-head attention followed by residual connections, LayerNorm, and a feed-forward network with an expansion ratio of four. The target-hypothesis branch retrieves alternative target hypotheses from the full temporal memory, the supportive branch retrieves evidence from the recent window of length $N_r = 28$, and the conflict branch retrieves inconsistent historical evidence from the full temporal memory. Padded memory positions are masked during attention.

The outputs of these mappings are followed by the normalization or gating operations defined in the main paper. Specifically, ϕ_α produces one scalar compatibility score for each target hypothesis, which is normalized across the N_Q hypotheses using Softmax. The scalar logit produced by ϕ_η is passed through a Sigmoid to obtain the conflict-calibration strength η_t. The four logits produced

Table S2: Architectures of the MLP-style mappings used in RTEM and TSAMF. $D = 256$ denotes the hidden feature dimension.

Block	Architecture
ϕ_e	Linear(D, D)–ReLU–Linear(D, D)
ϕ_q	Linear$(2D, D)$–ReLU–Linear(D, D)
ϕ_α	Linear$(3D, D)$–ReLU–Linear$(D, 1)$
ϕ_r	Linear$(3D, D)$–ReLU–Linear(D, D)
ϕ_s	Linear$(3D, D)$–ReLU–Linear(D, D)
ϕ_c	Linear$(4D, D)$–ReLU–Linear(D, D)
ϕ_η	Linear$(3D, D)$–ReLU–Linear$(D, 1)$
ϕ_ω	Linear$(4D, D)$–ReLU–Linear$(D, 4)$
ϕ_g	Linear$(3D, D)$–ReLU–Linear$(D, 1)$

by ϕ_ω are normalized using Softmax to weight the calibrated target representation $\tilde{r}_t$, supportive evidence $\mathcal{S}_t$, current observation h_t, and negative conflict evidence $-\mathcal{C}_t$. Finally, ϕ_g produces a scalar logit that is passed through a Sigmoid and scaled by $g_{\max}$ to obtain the current event gate g_t^e.

RTEM and TSAMF are learned end-to-end through the navigation objective without additional target-hypothesis-, support-, or conflict-level supervision. TSAMF refines only the current observation representation h_t, while the accumulated historical tokens are preserved.

A.2 Evaluation Protocol and Baseline Adaptation

Transient-Sound Evaluation Protocol. All methods are trained exclusively under the original target-sound condition. The transient-sound settings are introduced only at test time to evaluate robustness to reduced target-sound availability. Let the j-th target-sound event occupy an original active interval

$$\mathcal{I}_j = [t_j, t_j + L_j), \tag{11}$$

where t_j and L_j denote its onset and active duration, respectively. For a duration ratio $r \in (0, 1]$, the corresponding target-sound interval is constructed as

$$\mathcal{I}_j^{(r)} = [t_j, t_j + \max(1, \mathrm{round}(rL_j))). \tag{12}$$

Thus, as shown in Fig. S1, the onset times and number of target events remain unchanged across all conditions, while only the active duration of each target event is shortened and the removed portion is replaced by silence. The scene, agent start position, navigation goal, target-source location, interfering-source location, and interfering-source schedule are kept fixed across all conditions.

Baseline Configuration and Adaptation. Table S3 summarizes the configurations of the compared methods and their adaptation to the ENMuS^3 (Shi et al., 2025) multi-source navigation setting. We preserve the original model architecture and optimization strategy whenever applicable, while aligning the task interface, reward function, and multimodal observations with this common evaluation setting. Importantly, all methods are trained only on the original-sound training set; all transient duration ratios are introduced only at test time and are never used for training or fine-tuning.

Checkpoint Selection and Robustness Evaluation. For each method, a single checkpoint is selected according to the highest SPL on the validation set under the *original target-sound condition*. After selection, the model parameters are fixed and the same checkpoint is directly evaluated under all target-sound duration ratios reported in the main paper.

A.3 Mechanism Analysis

Fig. S2 analyzes the historical evidence retrieval behavior of RTEM from spatial, temporal, and cross-episode perspectives. Fig. S2(a) shows that the support and conflict branches exhibit distinct spatial retrieval patterns within the same representative successful episode: the former focuses more on target-related historical regions, whereas the latter assigns greater attention to regions that may contain competing acoustic evidence. Fig. S2(b) further shows that the two branches form different spatio-temporal attention patterns over accessible historical tokens, emphasizing evidence

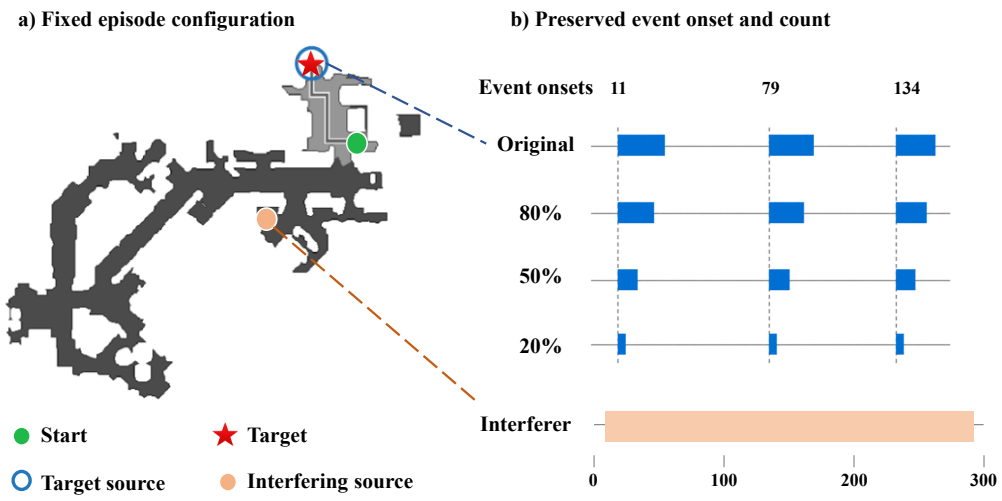


Figure S1: Episode-level construction of the transient-sound evaluation protocol. (a) The scene, start position, navigation goal, target source, and interfering source are fixed across all conditions. (b) Representative settings (Original, 80%, 50%, and 20%) preserve the onset and number of target events while progressively shortening their active durations. The same construction is applied to all evaluated duration ratios, while the interfering-source schedule remains unchanged.

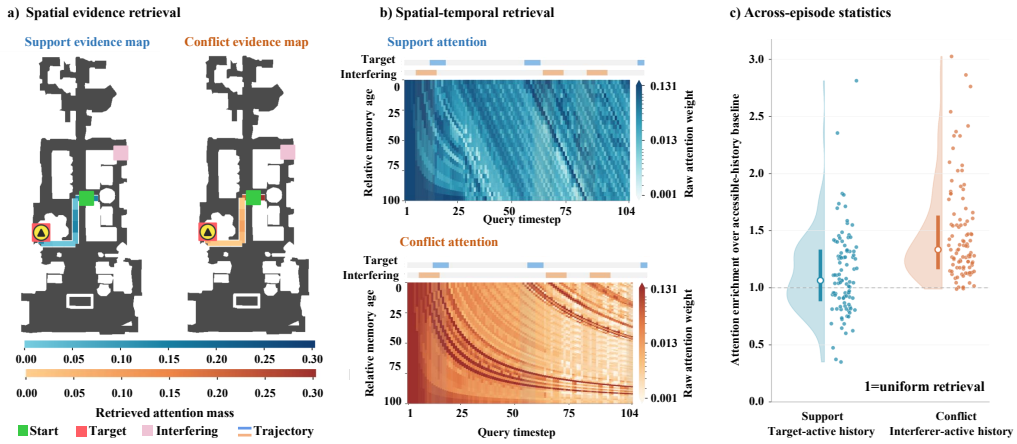


Figure S2: Mechanism analysis of RTEM. (a) Spatial retrieval densities of the support and conflict branches in a representative successful episode under the 20% target-sound condition. (b) Spatio-temporal retrieval patterns over accessible historical tokens, with target- and interferer-active intervals indicated above the attention maps. (c) Cross-episode attention enrichment over target-active and interferer-active history across 100 test episodes.

that reinforces the current target estimate and evidence useful for conflict calibration, respectively. Fig. S2(c) extends the analysis to 100 test episodes, where the support branch shows stronger attention enrichment over target-active history, while the conflict branch shows corresponding enrichment over interferer-active history. Overall, these results indicate a stable and complementary functional division between the two branches in historical evidence retrieval.

Fig. S3 illustrates how TSAMF modulates the contribution of event-level evidence over the navigation horizon. As valid historical context accumulates, the temporal modulation factor λ_t gradually

Table S3: Configurations and ENMuS^3 adaptations of the compared methods.

Method	Configuration	ENMuS^3 Adaptation
AV-Nav (Chen et al., 2020)	GRU hidden size 512; rollout length 150; 4 PPO epochs; 1 mini-batch; clip 0.1; learning rate 2.5×10^{-4}.	The original low-level action policy is retained and aligned with the common ENMuS^3 episodes, reward setting, and multi-source observation interface.
SAVi (Chen et al., 2021a)	Scene memory 150; hidden size 256; 8 attention heads; 1 encoder and 1 decoder layer; belief predictor; rollout length 150; 2 PPO epochs; 2 mini-batches; learning rate 2.5×10^{-4}.	The scene-memory and belief modeling components are retained and aligned with the common ENMuS^3 episodes and multi-source observations.
SMT+Audio (Fang et al., 2019)	Scene memory 150; hidden size 256; 8 attention heads; 1 encoder and 1 decoder layer; acoustic observations are incorporated into the SMT scene representation.	The audio-augmented SMT is evaluated with the common ENMuS^3 multimodal observations without introducing additional target information.
ENMuS^3 (Shi et al., 2025)	MSMT memory 159; hidden size 256; 8 attention heads; 1 encoder and 1 decoder layer; rollout length 160; 2 PPO epochs; 2 mini-batches; 32 environments; 60,000 updates.	The ENMuS^3 multi-source training setting is retained.
MAGNet (Zeng et al., 2026)	Two-stage training; scene memory 150; hidden size 256; 8 attention heads; GoalDescriptor memory 128; rollout length 150; 2 PPO epochs; 5 mini-batches; learning rate 2.5×10^{-4}.	The original GoalDescriptor, scene-memory architecture, and two-stage training are retained. We adapt only the task interface to the ENMuS^3 setting by aligning multimodal observations, action space, and reward settings. Transient-sound variants are evaluated without additional fine-tuning.
TraSNav	RTEM + TSAMF + MSMT; detailed architecture and optimization settings are reported in Table S1.	Trained on the same original-sound ENMuS^3 training split as the compared methods.

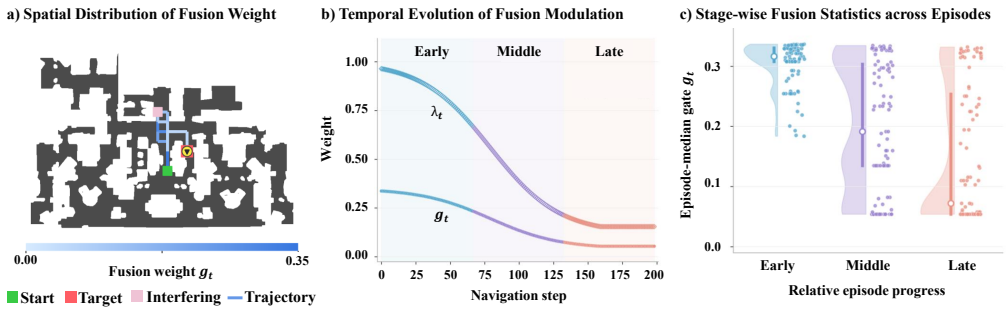


Figure S3: Mechanism analysis of TSAMF. (a) Final fusion weight g_t along a representative successful trajectory. (b) Temporal evolution of the modulation factor λ_t and final fusion weight g_t over the same episode. (c) Cross-episode distribution of the episode-median fusion weight over early, middle, and late navigation stages.

decreases and correspondingly attenuates the final fusion weight g_t. This gives event-level evidence a stronger contribution when long-term context is still limited, while progressively constraining its influence as scene and trajectory history becomes more complete.

Table S4: Training stability analysis of TraSNav over three independent runs.

Target Sound	SR ↑		SPL ↑	
	Mean	Std.	Mean	Std.
Original	48.96	0.14	26.90	0.11
80%	45.22	0.16	24.30	0.13
50%	38.37	0.24	19.67	0.20
20%	23.49	0.23	10.63	0.21

A.4 Training Stability Analysis

Since reinforcement learning-based navigation policies can be affected by training randomness, we further evaluate the stability of TraSNav across multiple independent training runs. The main experimental results in the paper follow a fixed random seed for fair comparison with existing methods, while this analysis reports the mean and standard deviation over three independent runs.

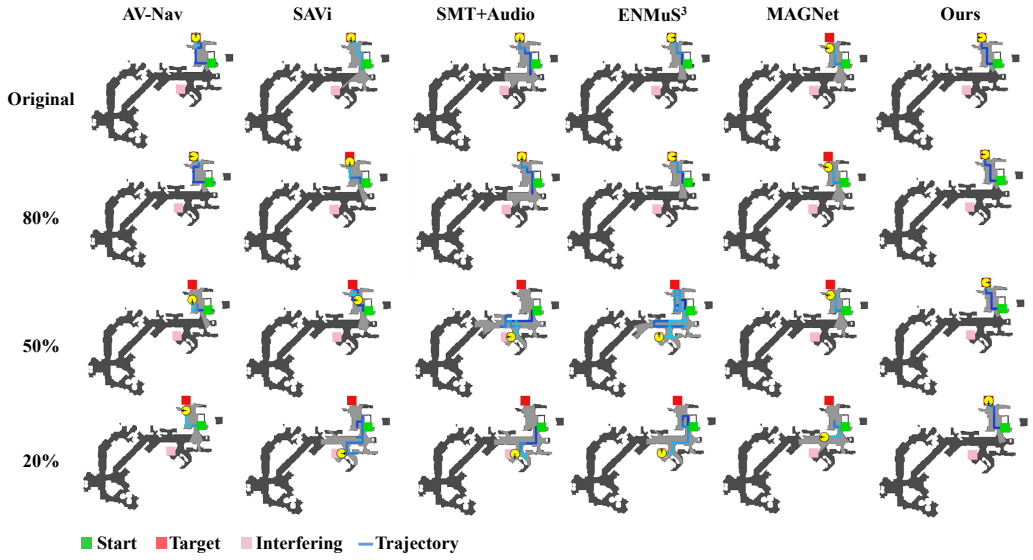
Table S4 summarizes the performance variation under different target-sound conditions. TraSNav maintains consistent navigation performance across different runs, with relatively small standard deviations on both success rate (SR) and success weighted by path length (SPL).

A.5 Qualitative Results

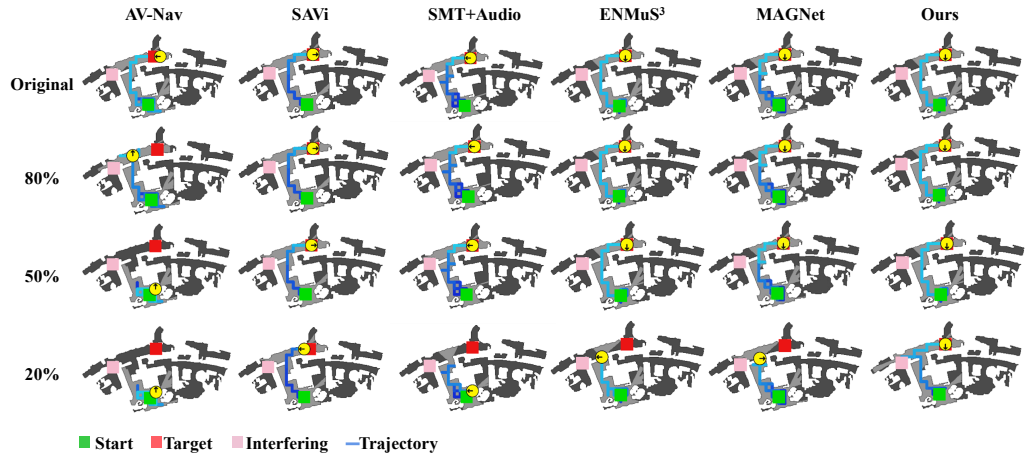
Fig. S4 provides additional episode-level trajectory comparisons under different target-sound durations. As the acoustic evidence available within each target event becomes increasingly limited, several baseline methods exhibit stronger local deviations, directional oscillations, or premature termination. This behavior suggests that, under sparse and potentially competing acoustic observations, individual short-term cues can exert unstable influence on navigation decisions. In contrast, TraSNav maintains more consistent goal-directed trajectories across different duration conditions, indicating that its decisions rely less on isolated acoustic observations and more on temporally verified target evidence.

A.6 Limitations and Future Work

TraSNav focuses on robustness to transient target sounds under the current target-conditioned navigation setting. While this provides a controlled basis for studying temporally sparse acoustic evidence, more general temporal event structures remain unexplored. A promising future direction is to move beyond predefined acoustic targets toward sound-triggered autonomous navigation, where an agent can determine whether an observed sound should initiate a navigation task and act accordingly in continuous soundscapes.



(a) Case 1



(b) Case 2

Figure S4: Additional episode-level trajectory comparisons among different audio-visual navigation methods under different target-sound durations. Each column corresponds to one navigation method, while each row denotes one target-sound condition (Original, 80%, 50%, and 20%). Green and red squares indicate the start and target positions, respectively; the pink square denotes the interfering source, the blue curve shows the executed navigation trajectory, and the yellow marker indicates the final agent position and heading.