D2AVN: EFFICIENT DIRECTION-DECOUPLED AUDIO-VISUAL NAVIGATION IN MULTI-SOURCE AND REVERBERANT SOUNDSCAPES

Anonymous authors
Paper under double-blind review

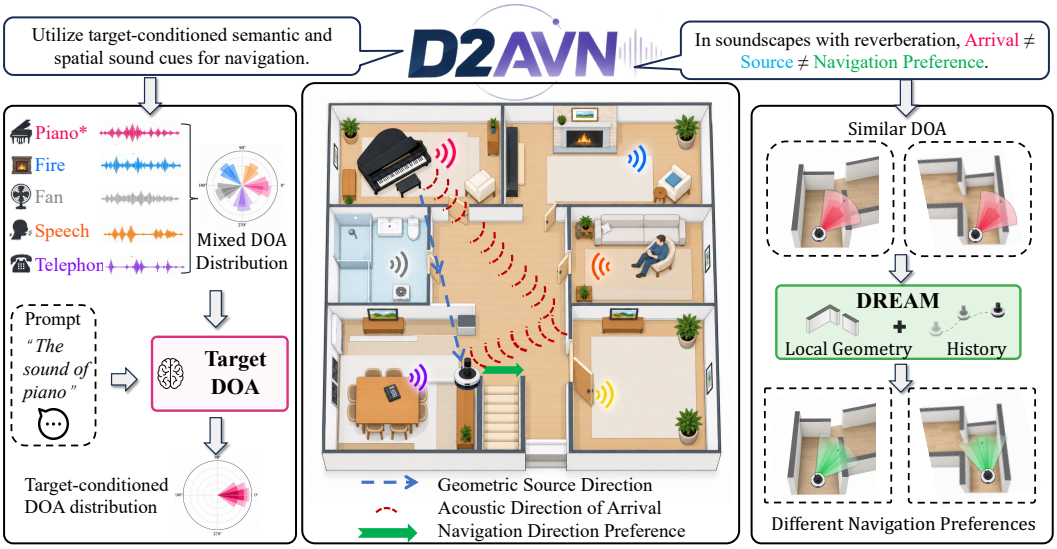

Figure 1: **Highlight of our proposed D2AVN framework.** D2AVN addresses audio-visual navigation in multi-source and reverberant environments by decoupling target-specific acoustic arrival direction from navigation-oriented directional preference. It isolates target arrival evidence from acoustic mixtures and directly adapts it into geometry-aware navigation preferences. **(a)** Target-DOA extracts target-specific acoustic arrival distributions from multi-source mixtures, conditioned on a target-text query. **(b)** geometric source direction (blue), acoustic Direction of Arrival (DOA) (pink), and navigation direction preference (green) can differ due to reverberation and local geometry. **(c)** DREAM adapts arrival evidence using local geometry and egomotion-aligned history, allowing similar acoustic arrival cues to yield different navigation preferences.

ABSTRACT

Audio-Visual Navigation (AVN) empowers embodied agents to leverage the target-specific acoustic cues to complement visual perception, enabling them to locate sound sources that are visually occluded or outside their immediate field of view. Previous AVN methods predominantly derive directional guidance by estimating the *geometric source direction*—the direct spatial vector from the agent to the target. However, this paradigm struggles in authentic indoor environments, where multi-source interference obscures target cues and reverberation causes the perceived acoustic *Direction of Arrival (DOA)* to diverge significantly from the direct source path. To this end, we propose **D2AVN**, a novel *direction-decoupled* framework that explicitly isolates target acoustic arrival evidence before translating it into actionable navigation guidance. Importantly, our D2AVN employs a novel target-conditioned DOA (**TargetDOA**) module that extracts target-specific arrival distributions directly from complex acoustic mixtures, utilizing paired target-only acoustic references that exclude competing sources while preserving environmental propagation. Then, our Depth-conditioned Reward-Aligned Ego-

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.

centric Affordance Modeling (**DREAM**) module reshapes these arrival distributions into reward-aligned navigation direction preferences by incorporating local geometry and egomotion-aligned history. To facilitate robust policy learning in these complex spaces, we implement a training pipeline with effective online acoustic rendering and reinforcement learning. Extensive validation demonstrates that D2AVN achieves a **33.0%** success rate with First-Order Ambisonics (FOA) and **15.8%** with binaural audio. Notably, the binaural configuration improves Success weighted by Path Length (SPL) and action efficiency (SNA) by **74%** and **90%**, respectively, over the fine-tuned ENMuS^3 baseline.

1 INTRODUCTION

Navigation requires embodied agents to reach specified goals under partial observations in unseen 3D environments (Anderson et al., 2018; Savva et al., 2019; Gao et al., 2026; Pan et al., 2026a). Many methods rely primarily on egocentric visual observations, which are constrained by limited fields of view and scene occlusions (Mirowski et al., 2016; Zhu et al., 2017; Gupta et al., 2017; Wijmans et al., 2019). Acoustic observations carry both semantic and spatial information, providing complementary target cues when direct visual evidence is unavailable. Audio-Visual Navigation (AVN) leverages acoustic cues to complement visual perception, enabling agents to navigate toward sound-emitting targets that are visually occluded or outside their current field of view (Chen et al., 2020; 2021a; 2022).

Existing AVN methods incorporate acoustic guidance through 1) implicit policy representations and 2) explicit localization. The former encodes spectrograms or acoustic maps into latent features, which are fused with visual features to learn a navigation policy through reinforcement learning, maximizing cumulative navigation rewards (Chen et al., 2020; 2021b). However, these representations may entangle semantic and spatial information, increasing sensitivity to sound content and competing sources (Wang et al., 2023; Shi et al., 2025b). The latter instead estimates the *geometric source direction*—the direction from the agent to the target—or its relative position, providing interpretable directional guidance. This interpretable spatial formulation offers a practical basis for navigation guidance and has been adopted in a range of AVN methods (Gan et al., 2020; Shi et al., 2025b; Zeng et al., 2026).

Research Problem. However, obtaining reliable guidance through explicit localization remains challenging in multi-source and reverberant indoor environments (Lee et al., 2026; Opochinsky et al., 2025). Competing sounds obscure target-specific directional cues (Jia et al., 2021; Li et al., 2017), while reverberation and occlusion can cause the acoustic *Direction of Arrival (DOA)* (Gerstoft et al., 2016; Zheng et al., 2024) to differ from geometric source direction (Fig. 1(b)). DOA describes the direction from which target sound reaches the agent and therefore reflects environmental propagation (An et al., 2018; Chen et al., 2022). Together, *multi-source interference and propagation-dependent arrivals make the geometric source direction difficult to infer reliably from acoustic observations*. Moreover, *navigating towards the target requires more than obtaining a reliable target-conditioned DOA*. The agent must also account for local geometry when choosing its movements: obstacles may require turns or detours rather than movement along the acoustic arrival direction (Gan et al., 2020; Chen et al., 2021b). A navigation direction preference therefore reflects which directions are useful for navigating toward the target under these geometric constraints and need not coincide with the DOA. These distinctions motivate an interesting idea of *separating target-specific acoustic perception from navigation-oriented directional adaptation*.

Our Contributions. In light of this, we propose **D2AVN**, a novel *direction-decoupled* framework that isolates target acoustic arrival evidence before translating it into actionable navigation guidance. Our D2AVN brings two key technical contributions. Firstly, we introduce Target-conditioned DOA (**TargetDOA**) that estimates target-specific arrival distributions from acoustic mixtures conditioned on a target query (Fig. 1(a), Sec. 3.2). Instead of geometric source direction, its supervision uses paired target-only DOA that exclude competing sources while preserving environmental propagation. On top, we then propose Depth-conditioned Reward-Aligned Egocentric Affordance Modeling (**DREAM**). It transforms these distributions into reward-aligned navigation direction preferences using local depth geometry and egomotion-aligned history (Fig. 1(c), Sec. 3.3). Navigation rewards guide this transformation through the downstream actor-side objective, allowing DREAM to learn how to incorporate geometry and history into directional adaptation. The resulting navigation pref-

erences condition RGB-D fusion for recurrent action prediction, with the policy optimized using PPO while TargetDOA remains frozen.

To support training and evaluation in multi-source and reverberant acoustic conditions, we construct a custom simulation setting by segmenting long-duration audio recordings into source clips and integrating them into SoundSpaces 2.0 (Chen et al., 2022) (Sec. 3.5). Online rendering of multiple concurrent sources incorporates environment-dependent sound propagation, producing soundscapes designed to better reflect real-world indoor acoustic conditions. Our default configuration uses First-Order Ambisonics (FOA) (Au et al., 2021), while a binaural variant enables input-modality-matched comparisons with binaural baselines. Extensive evaluation in this setting demonstrates that D2AVN achieves success rates of **33.0%** with FOA and **15.8%** with binaural audio. Notably, the binaural configuration achieves relative improvements of **74%** in Success weighted by Path Length (SPL) and **90%** in action efficiency (SNA) (Chen et al., 2021a) over the fine-tuned ENMuS^3 baseline.

2 Related Works

Audio-Visual Navigation. Audio-visual navigation combines auditory and visual observations to guide embodied agents toward sounding targets. Early work integrates visual mapping (Konolige & Agrawal, 2008), sound-source localization (Risoud et al., 2018), and path planning (Gan et al., 2020), while SoundSpaces establishes end-to-end reinforcement learning from egocentric audio-visual observations in acoustically rendered 3D environments (Chen et al., 2020; 2021b; Younes et al., 2023). Semantic Audio-Visual Navigation further associates acoustic events with semantic objects and introduces short-duration sounds, motivating persistent multimodal memory for navigation after the target becomes silent (Chen et al., 2021a). More recent studies address interference and temporal continuity. ENMuS^3 extracts target-specific semantic and spatial features from multi-source binaural observations and maintains multi-scale scene memory for navigation in noisy environments (Shi et al., 2025b). SAVN-CE introduces spatially and temporally continuous audio-visual observations, with MAGNet combining historical context and self-motion to maintain goal information during acoustic absence (Zeng et al., 2026). These approaches advance target inference and temporal reasoning, but leave the relationship between acoustic arrival evidence and navigation-effective directional guidance implicit. *Different from these works, our goal is to explicitly decouple target-specific acoustic perception from navigation-oriented directional adaptation.*

Task-Oriented Spatial Representations for Navigation. Intermediate spatial representations connect multimodal perception with navigation decisions (Hong et al., 2023; Zhang et al., 2025). AV-WaN combines spatially grounded acoustic memory with dynamically predicted waypoints learned end-to-end through reinforcement learning (Chen et al., 2021b). Goal-descriptor approaches instead maintain semantic and spatial target information within persistent representations (Chen et al., 2021a; Zeng et al., 2026). Other methods regularize spatial features or adapt cross-modal fusion. RAVN trains an acoustic geometry reasoner with geometric proxy supervision and uses its learned reliability cues to modulate visual features (Liu & Yu, 2026). BDATP explicitly models binaural differences and introduces action-transition prediction as an auxiliary objective for generalizable navigation representations (Li & Yu, 2026). These formulations provide spatial memory, goal descriptors, reliability-aware features, or navigation subgoals, but do not explicitly formulate directional guidance as a transformation of target-conditioned acoustic arrival evidence. *By contrast, our goal is to transform target-specific acoustic evidence into an explicit, reward-aligned directional prior. As part of the policy observation, this transformed directional bias helps the RL agent exploit DOA cues more effectively.*

Target-Conditioned Acoustic Perception and Localization. Sound event localization and detection jointly identifies acoustic events and estimates their spatial directions (Adavanne et al., 2018; Cao et al., 2019). SELDnet learns event activities and class-associated directions from multichannel recordings, while ACCDOA unifies activity and direction estimation through activity-coupled Cartesian vectors (Adavanne et al., 2018; Shimada et al., 2021). Beyond predefined event classes, text-queried sound event localization uses natural-language descriptions to specify the target source (Zhao et al., 2024). LuSeeL jointly performs language-conditioned binaural sound extraction and localization (Pan et al., 2026b), while SelectTSL uses prompt-guided spatial features to estimate target directions and the number of active target sources (Jiang et al., 2026). Although these methods improve target selectivity, their localization objectives recover annotated source directions. In particular, the directional distributions in text-queried localization and LuSeeL are constructed around

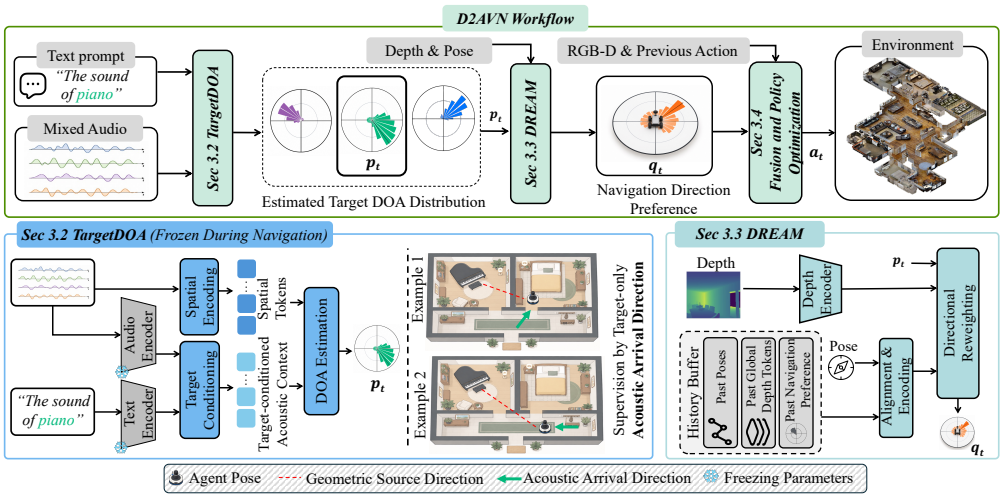

Figure 2: **Overview of D2AVN.** TargetDOA extracts target-specific arrival evidence p_t, with supervision derived from each mixture's target-only acoustic component (Sec. 3.2). DREAM adapts p_t into a navigation preference q_t using current depth geometry and pose-aligned history (Sec. 3.3). This preference conditions RGB-D fusion for recurrent action prediction (Sec. 3.4). TargetDOA remains frozen during navigation learning.

reference source azimuths, rather than derived from the target's reverberant arrival evidence (Zhao et al., 2024; Pan et al., 2026b). *In contrast, our goal is to extract target-specific acoustic arrival evidence for navigation in multi-source and reverberant environments. This disentangles target cues from competing sounds while preserving reflection-dependent directional information.*

3 The Proposed D2AVN Framework

3.1 Overview

We consider target-conditioned audio-visual navigation in multi-source, reverberant indoor environments. Given a text query c, the agent navigates to the corresponding sound source and declares arrival. At step t, it receives RGB I_t, metric depth D_t, one-second FOA audio $X_t \in \mathbb{R}^{4\times 16000}$, pose ξ_t, and self-state u_t, and selects among MoveForward, TurnLeft, TurnRight, and Stop.

D2AVN separates target-specific acoustic perception from navigation adaptation and integrates directional guidance with vision. As shown in Fig. 2, **TargetDOA** first estimates a target-specific acoustic arrival distribution p_t from X_t and c, following pretraining with paired target-only acoustic references rather than geometric source bearings. **DREAM** then adapts p_t into a navigation-preference distribution q_t using local depth geometry and yaw-aligned, pose-conditioned history. This adaptation is learned through the downstream actor-side objective with TargetDOA frozen. Finally, q_t conditions RGB-D fusion in a shared egocentric polar representation and is explicitly retained in the fused state for recurrent action prediction. Both $p_t, q_t \in \Delta^{K-1}$ span $K = 24$ uniformly spaced egocentric azimuth sectors, with forward at zero azimuth and positive angles to the right.

3.2 TargetDOA: Target-Conditioned Acoustic Arrival Estimation

Insight: Isolate target evidence from mixtures while preserving reverberation-dependent directional cues. **TargetDOA** estimates the queried target's arrival distribution using target-conditioned semantics and spatial acoustic features. Both variants use target-only acoustic references rather than geometric bearings. The default model uses four-channel FOA input and $K = 24$ egocentric azimuth bins. For input-modality-matched comparisons with binaural baselines, we evaluate a two-channel, 12-bin variant motivated by front–back ambiguity. Unless otherwise stated, the formulations below describe the FOA configuration.

Target-only acoustic supervision. We render multi-source audio mixtures with SoundSpaces 2.0 (Chen et al., 2022) and retain their per-source components. Each mixture is paired with two dis-

tinct in-scene queries and their corresponding target-only references, encouraging query-dependent predictions from the same waveform. Each reference shares the mixture's scene, receiver pose, and observation window, preserving target propagation while excluding competing sources.

For FOA, a window-averaged horizontal active-intensity proxy defines the egocentric reference under our world-frame convention:

$$\vartheta_t^{\mathrm{aco}} = \mathrm{wrap}[\mathrm{atan2}(\langle WZ\rangle_t, \langle WX\rangle_t) - \psi_t]\,, \tag{1}$$

where W, X, Z are channels of the target-only component for query c, $\langle UV\rangle_t$ is the sample average of $U_n V_n$ over the observation window, and ψ_t is the receiver heading. This reference summarizes the target's net acoustic direction, including propagation effects, without resolving individual paths. For sector centers θ_k, normalized soft labels are $y_t(k) \propto \exp[8\cos(\theta_k - \vartheta_t^{\mathrm{aco}})]$. Query dependence is implicit in p_t, y_t, and $\vartheta_t^{\mathrm{aco}}$.

Target-conditioned DOA estimation. As shown in Fig. 2, the frozen *Text Encoder* (T5-base) and *Audio Encoder* (DAC-VAE) from SAM Audio (Shi et al., 2025a) produce text tokens T_c and audio tokens A_t, respectively. *Spatial Encoding* extracts modality-specific acoustic features. For FOA, horizontal components are rotated into the receiver frame using ψ_t, then all four channels are scaled by a shared window-level RMS normalization gain. The Audio Encoder receives the mixture's omnidirectional W channel without this normalization. Time–frequency intensity, coherence, diffuseness, and energy features are pooled across frequency using learned spatial reliability and relative signal energy, then temporally encoded into spatial tokens S_t. In *Target Conditioning*, text tokens query audio tokens through cross-attention to form a target-conditioned acoustic context. In *DOA Estimation*, spatial tokens query this context through a second cross-attention block, yielding target-conditioned spatial tokens $\widetilde{S}_t$. Temporal pooling combines learned token scores, spatial reliability, and energy to obtain h_t^{aco}. A two-layer MLP F_{doa} with a GELU hidden activation predicts direction logits, followed by softmax:

$$p_t = \mathrm{softmax}(F_{\mathrm{doa}}(h_t^{\mathrm{aco}}))\,. \tag{2}$$

Acoustic pretraining. For the FOA configuration, we minimize soft-label cross-entropy plus a directional cosine loss over mixture windows and paired queries:

$$\mathcal{L}_{\mathrm{aco}} = \mathbb{E}_{t,c}\left[-\sum_{k=0}^{K-1} y_t(k)\log p_t(k) + 1 - \widehat{\mu}_t^{\top} e(\vartheta_t^{\mathrm{aco}})\right], \tag{3}$$

where $e(\theta) = [\sin\theta, \cos\theta]^{\top}$, $\mu_t = \sum_k p_t(k) e(\theta_k)$, and $\widehat{\mu}_t = \mu_t / \max(\|\mu_t\|_2, 10^{-6})$.

Here, FOA pretraining applies one random virtual-heading rotation to each mixture group, consistently transforming its horizontal components and references across both queries and all group samples. Meanwhile, evaluation retains deterministic preprocessing but disables random augmentation. Target-only references are used solely for pretraining supervision. Moreover, inference uses the mixed audio and query, with heading used for FOA coordinate conversion. The acoustic front end is frozen during navigation learning. The FOA implementation details and the binaural variant's input processing, reference construction, angular encoding, training objective, and interface to DREAM are provided in Appendix A.

3.3 DREAM: Depth-conditioned Reward-aligned Egocentric Affordance Modeling

Insight: Acoustic arrival evidence need not indicate an effective movement direction. **DREAM** adapts p_t using current depth geometry and pose-conditioned history, learning q_t through the actor-side navigation objective rather than geometric-bearing or oracle direction supervision. The output retains the angular representation of p_t but is not another source-localization estimate.

Current geometry encoding. As shown in Fig.2, the *Depth Encoder* extracts patch tokens $\mathbf{D}_t$ and a global token d_t^{g} from D_t. Within *Directional Reweighting*, a directional encoder maps p_t and its distribution-derived features to K tokens, preserving circular adjacency across the azimuth boundary. These tokens query all current depth patches through cross-attention, yielding one geometry-conditioned context token per sector. The global token is retained for history updates.

History alignment and encoding. The *History Buffer* $\mathcal{B}_t$ stores up to 16 past states $(q_s, d_s^{\mathrm{g}}, \xi_s, v_s)$, comprising navigation preferences, global depth contexts, receiver poses, and validity flags, with $s < t$. Using the current pose, *Alignment & Encoding* applies a fractional circular shift $\mathcal{R}_{s\to t}$ based on relative yaw to align historical distributions with the current angular frame. Relative translation is a conditioning feature, not an exact geometric reprojection. Aligned distributions, depth contexts, relative poses, and ages form history tokens $\widetilde{\mathcal{B}}_t$. Within Directional Reweighting, the current geometry-conditioned tokens query this history through temporal cross-attention to produce sector contexts $\mathbf{H}_t$. History also provides an explicit prior:

$$m_t = \mathcal{N}\left(\sum_{s\in\mathcal{B}_t} v_s e^{-\lambda_{\mathrm{age}} a_{s,t}} \mathcal{R}_{s\to t}(q_s)\right), \tag{4}$$

where v_s indicates valid directional history, $a_{s,t}$ is its age, and λ_{age} controls decay. $\mathcal{N}$ normalizes mass, returning a uniform distribution when no valid history exists. History thus contributes both contextual features and a distribution-level reference. The current tuple is appended only after prediction; the buffer resets at episode boundaries.

Directional reweighting. From pooled $\mathbf{H}_t$, distribution statistics, and history availability, DREAM predicts routing and correction gates $\alpha_t, g_t \in [0,1]$. A sector-wise head produces $r_t \in \mathbb{R}^K$ with $\sum_k r_t(k) = 0$ and $|r_t(k)| < 2$. Current and historical evidence are mixed before gated log-space correction:

$$p_t^{\mathrm{base}} = \mathcal{N}(\alpha_t p_t + (1-\alpha_t) m_t), \tag{5a}$$

$$q_t = \mathrm{softmax}\left(\log_\epsilon p_t^{\mathrm{base}} + \frac{g_t r_t}{\tau}\right). \tag{5b}$$

Here $\log_\epsilon x = \log\max(x,\epsilon)$ is elementwise, and $\tau > 0$ scales the residual. When only current or historical evidence is valid, α_t is fixed to 1 or 0, respectively. The routing gate balances current and historical evidence, while the gated residual adapts the preference to geometry and history. Validity, gate, residual, and state-update details appear in Appendix B.

3.4 Fusion and Policy Optimization

Preference-conditioned fusion. A trainable RGB encoder extracts appearance features, which are aligned with depth-derived geometry through calibrated projection and interpolation on a $K \times R$ egocentric polar grid, with $R = 5$ radial bins. The grid shares the angular convention of q_t and uses current RGB-D observations. Sector tokens derived from q_t condition radial fusion, followed by neighboring-sector context exchange and bounded preference-based modulation. Polar aggregation yields $z_t^{\mathrm{fuse}} = [q_t; z_t^{\mathrm{tail}}] \in \mathbb{R}^{512}$, where z_t^{tail} summarizes fused scene context and self-state u_t. The explicit concatenation retains direct policy access to q_t.

Recurrent policy and optimization. Separate policy and value GRUs both receive the previous action a_{t-1}. The policy GRU receives z_t^{fuse}, whereas the value GRU receives $\mathrm{sg}(z_t^{\mathrm{fuse}})$. Their actor and critic heads predict the action distribution π_t and state value V_t, respectively. Stop-gradient sg prevents value-regression gradients from updating DREAM or the fusion network. We minimize an actor-side objective combining clipped PPO (Schulman et al., 2017), an action-entropy bonus, and a penalty on update-induced changes in navigation preference:

$$\begin{aligned}\mathcal{L}_{\mathrm{policy}} = \mathcal{L}_{\mathrm{PPO}} &- \lambda_H(S)\mathbb{E}_t[\mathcal{H}(\pi_t)] \\ &+ 0.01\,\mathbb{E}_t\left[D_{\mathrm{KL}}\left(\mathrm{sg}(q_t^{\mathrm{old}}) \| q_t^{\mathrm{new}}\right)\right],\end{aligned} \tag{6}$$

where $\mathcal{L}_{\mathrm{PPO}}$ is the negative clipped surrogate objective, and $\lambda_H(S)$ follows the stage-specific schedules in Appendix C.2. The KL term anchors update-time q_t^{new} to its fixed rollout-time reference q_t^{old}. During recurrent updates, q_t^{new} is recomputed causally from detached sequence-entry states. The value branch uses clipped regression with return targets constructed from generalized advantage estimation (GAE). Rewards combine geodesic progress, a per-step cost, a successful-stop bonus, and penalties for timeouts and repeated turning. Fusion and training details are provided in Appendix C.

3.5 Custom Simulation Environment

We construct a multi-source, reverberant simulation environment for training and evaluation in Matterport3D (Chang et al., 2017) using Habitat-Sim (Savva et al., 2019) and SoundSpaces 2.0 (Chen et al., 2022). Each episode contains one target and two to seven background emitters, drawn from 14 target-sound classes and 10 background-sound classes, respectively. Emitters remain stationary within each episode, while source–receiver room impulse responses (RIRs) are rendered online as the agent moves or turns. Rendering incorporates direct and indirect propagation, diffraction, and transmission. Target audio cycles through ordered, pre-segmented clips, while background clips are configured and looped separately, with repeated classes or recordings permitted. Each dry-audio window is normalized before source-wise convolution and mixing, producing 1-s observations at 16 kHz with four-channel FOA or binaural receivers. Further details are provided in Appendix D.

4 Experiments

4.1 Comparison with Baselines

Evaluation protocol and metrics. Methods share evaluation episodes and success criteria within each protocol. Checkpoints for trained or adapted models are selected on validation data. Navigation-training and evaluation scenes are disjoint. We report success rate (SR), path- and action-weighted success (SPL and SNA), and final distance to goal (DTG). SR is reported in percent, DTG is in meters, and the remaining metrics lie in $[0, 1]$. All methods have a budget of 500 primitive actions and succeed only by issuing Stop in a designated range from the target. All failures and timeouts are included in the reported metrics.

Single-source transfer. We evaluate frozen policies on 994 fixed episodes from the SAVi MP3D semantic-navigation test split, excluding six episodes with target categories unsupported by ENMuS^3. Each episode uses default SoundSpaces 2.0 propagation and a single source looping its full waveform, omitting the original emission duration and start offset. No method receives additional training or fine-tuning. Baselines include Random, ObjectGoal (Batra et al., 2020; Wijmans et al., 2019), AV-Nav (Chen et al., 2020), AV-WaN (Chen et al., 2021b), and ENMuS^3 (Shi et al., 2025b). ENMuS^3 uses its released checkpoint; AV-WaN retains its waypoint controller, with all primitive actions counted toward the budget and SNA. Table 1 shows that D2AVN achieves the highest SR, SPL, and SNA among the evaluated methods. Compared with ENMuS^3, its modest SR gain is accompanied by substantially higher SPL and SNA, highlighting success-weighted path- and action-efficiency benefits under single-source transfer.

Table 1: **Frozen-policy transfer on SAVi-derived single-source episodes.** Best values are bold.

Method	SR↑	SPL↑	SNA↑	DTG↓
Random	0.10	0.0010	0.0004	18.03
ObjectGoal	2.62	0.0082	0.0028	17.39
AV-Nav	3.32	0.0065	0.0025	11.12
AV-WaN	6.14	0.0154	0.0112	15.24
ENMuS^3	19.22	0.05121	0.03069	**5.74**
D2AVN(Ours)	**20.12**	**0.1039**	**0.0592**	6.86

Dense multi-source navigation. We compare ENMuS^3 with both D2AVN variants in the same custom simulation environment, using common navigation training splits, source configurations, playback rules, target information, action spaces, and rewards.

Tab. 2 reports complete evaluations on the same fixed set of 500 held-out episodes from unseen scenes, separating the binaural comparison from the FOA configuration. In the binaural comparison, D2AVN improves SR over ENMuS^3 and achieves relative improvements of 74.4% in SPL and 89.8% in SNA. D2AVN–FOA achieves the strongest results across all four reported metrics; this configuration is evaluated separately as it uses a different audio input. Figure 3 shows D2AVN successfully reaching the target amid multiple distractors where ENMuS^3 fails, providing qualitative support for the observed gains in success-weighted navigation efficiency.

Table 2: **Multi-source navigation.**

Method	Audio	SR↑	SPL↑	SNA↑	DTG↓
ENMuS^3	Binaural	12.4	0.0655	0.0361	**11.99**
Ours	Binaural	**15.8**	**0.1142**	**0.0685**	12.67
Ours	FOA	33.0	0.1610	0.0868	9.95

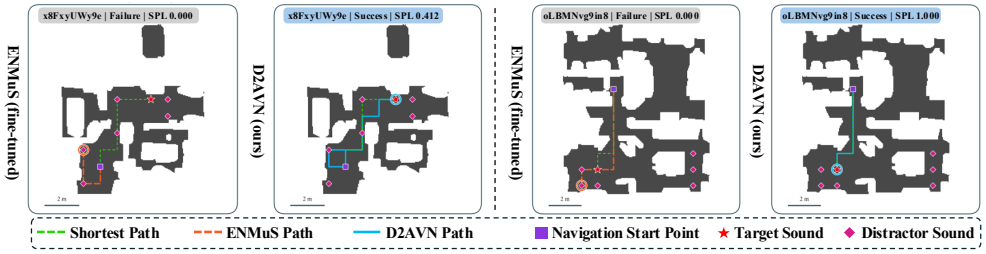

Figure 3: **Qualitative navigation comparison.** Each pair shows fine-tuned ENMuS[3] (left) and D2AVN (right) in the same scene. Orange dashed and blue solid lines denote their respective trajectories; green dashed lines indicate shortest paths.

4.2 Component Analysis

TargetDOA: semantic, spatial, and supervision controls. We analyze TargetDOA with three controlled variants. *Without target prompt* replaces the query with a shared learned token; *Without spatial cues* zeros the directional FOA channels while retaining the spatial encoder; and *Geometric supervision* replaces the target-only acoustic reference with the source bearing. All variants use the same learned 24-bin direction head and matched training settings. We evaluate three seeds on the same 240 mixtures from 15 scenes, querying both targets on each waveform.

Table 3 shows that removing either the target prompt or spatial cues substantially degrades acoustic-direction estimation, indicating that semantic target selection and spatial acoustic evidence both contribute to target-specific arrival estimation. Geometric supervision improves geometric-bearing accuracy but performs worse on the acoustic reference, supporting the distinction between propagation-dependent arrival estimation and geometric source localization. Although the prompt-free variant attains slightly lower CE, its much poorer angular metrics suggest that CE alone does not fully capture target-selective directional accuracy. Results are macro-averaged across scenes, mixtures, and seeds. In Table 3, A/G denote acoustic/geometric supervision; MAE is in degrees, A30 is acoustic-reference Accuracy@30° in percent, and CE uses acoustic soft labels with $\kappa = 8$.

Table 3: **TargetDOA controls.**

Variant	Ref.	$\text{MAE}_{\text{aco}} \downarrow$	A30↑	$\text{MAE}_{\text{geo}} \downarrow$	CE↓
Full TargetDOA	A	**37.74**	**64.72**	50.81	2.961
Without target prompt	A	60.69	28.75	66.47	**2.885**
Without spatial cues	A	91.99	15.28	89.10	3.203
Geometric supervision	G	50.13	51.53	**39.27**	3.625

Table 4: **DREAM interventions.**

Variant	SR↑	SPL↑	SNA↑	DTG↓
Full DREAM	**33.0**	**0.1610**	**0.0868**	**9.95**
Residual-off	29.2	0.1534	0.0854	11.03
No-history	32.2	0.1466	0.0783	10.10
TargetDOA-direct	28.6	0.1534	0.0857	10.92

DREAM: from acoustic evidence to navigation preference. We evaluate three inference-time interventions on one trained checkpoint over 500 fixed episodes in unseen scenes. *Residual-off* sets $g_t = 0$ while retaining history and evidence routing; *No-history* disables DREAM history while preserving current-observation processing; and *TargetDOA-direct* bypasses DREAM by supplying $q_t = p_t$. All variants share weights and episode specifications but follow their own closed-loop trajectories, thereby assessing how the trained policy uses DREAM rather than comparing independently retrained architectures.

Full DREAM achieves the best aggregate performance across all reported metrics (Tab. 4), improving SR from **28.6%** to **33.0%** over Dream off. Residual-off causes a larger SR drop than No-history, while removing history more strongly reduces SPL and SNA, suggesting complementary roles for directional correction and temporal context. These results support adapting acoustic arrival evidence into navigation preferences within the trained policy. Figure 4 illustrates this adaptation at the action level: by incorporating local geometry and history, DREAM shifts the policy's action probabilities at the marked poses, favoring decisions that support more effective navigation in the shown examples.

We further examine performance across initial shortest-path distance. As shown in Fig. 5, Full DREAM improves mid-range SR from **28.0%** to **38.0%** relative to TargetDOA-direct and increases long-range successes. Although No-history matches Full DREAM in mid-range SR, retaining history yields higher SPL and SNA, indicating that temporal context mainly improves navigation efficiency. The gains are not uniform across all distances: Full DREAM is weaker at short range but shows greater benefits in mid- and long-range navigation.

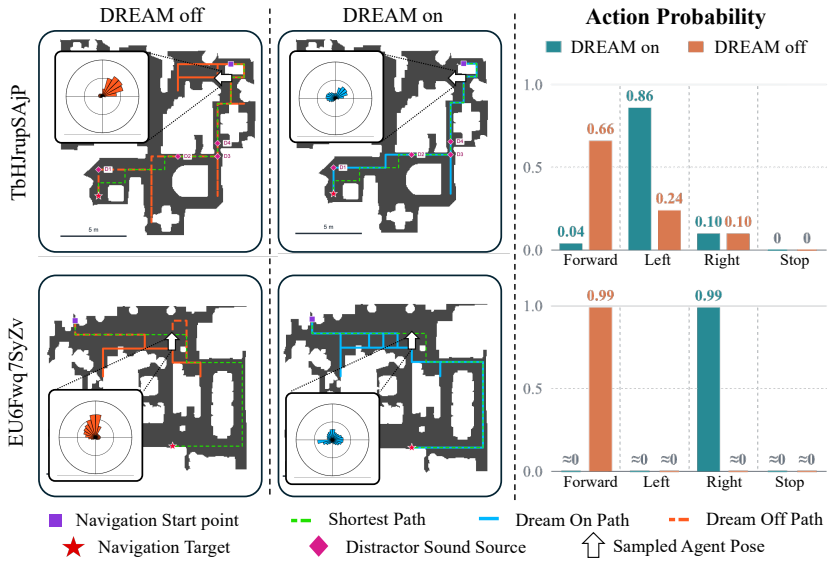

Figure 4: **DREAM adapts directional guidance for geometry-aware navigation.** Orange and blue denote *DREAM off* and *DREAM on* trajectories, respectively; green dashed lines show shortest paths. Polar insets show the directional distributions forwarded downstream at the marked poses; bar plots show action probabilities.



Figure 5: **DREAM interventions on mid- and long-range navigation.** Mid-range and Long-range denote initial shortest navigable path lengths of $[10, 20)$ m and $[20, 50)$ m, respectively.

5 Conclusion, Limitations and Future Work

Conclusion. We presented **D2AVN**, a direction-decoupled framework for audio-visual navigation in multi-source and reverberant environments. Specifically, TargetDOA extracts target-conditioned arrival evidence from mixtures using propagation-preserving target-only acoustic references, while DREAM adapts this evidence into reward-aligned navigation preferences using local geometry and egomotion-aligned history. These preferences, in turn, condition visual–geometric fusion and recurrent action prediction. Experiments show improved SR, SPL, and SNA over ENMuS^3 in the binaural comparison, while component analyses support the roles of target conditioning, spatial acoustic cues, and directional adaptation. Together, these results highlight direction decoupling as a means of translating propagation-dependent acoustic evidence into geometry-aware guidance for more efficient navigation.

Limitations and Future Work. Our simulation-based evaluation does not fully capture real microphone responses or room acoustics. Moreover, the multi-source protocol incorporates target-level safeguards to preserve audibility. Reliable navigation under near-complete target masking is not ensured, and the minimum required target audibility remains unquantified. Accordingly, future work will investigate real-world transfer, moving sources, and broader acoustic conditions, including navigation robustness across controlled levels of target audibility.

AI USE STATEMENT

In this work, we used generative AI tools to obtain feedback on research methodology and experiments and to assist with interpreting results. Additionally, we used these tools for manuscript drafting and language polishing, literature retrieval and discovery, and creating schematic illustrations and graphical elements for figures. We have reviewed all AI-assisted work. Specifically, we checked methodological descriptions and interpretations against our implementation and experimental records, verified cited references against their original sources, and inspected generated illustrations for consistency with their intended schematic meaning. The generated illustrations serve explanatory purposes rather than constituting experimental evidence. We take responsibility for the final content of this work, including text, claims, references, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies audio-visual navigation in simulated indoor environments using existing scene, audio, and pretrained-model resources. The reported navigation experiments involve simulated agents. Our simulation-based evaluation does not establish deployment safety, which would require additional validation.

REPRODUCIBILITY STATEMENT

Section 3 describes the model architecture, acoustic supervision, and navigation optimization, while Section 4 specifies the evaluation protocols, baseline configurations, metrics, and component studies. Additional implementation and training details for TargetDOA and DREAM are provided in the appendices. The custom environment's rendering, playback, and mixing settings are detailed in Appendix D.

REFERENCES

Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018.

Inkyu An, Myungbae Son, Dinesh Manocha, and Sung-Eui Yoon. Reflection-aware sound source localization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 66–73. IEEE, 2018.

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

Eugena Au, Shirley Xiao, CT Justine Hui, Yusuke Hioka, Hinako Masuda, and Catherine I Watson. Speech intelligibility in noise with varying spatial acoustics under ambisonics-based sound reproduction system. *Applied Acoustics*, 174:107707, 2021.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020.

Yin Cao, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, and Mark D Plumbley. Polyphonic sound event detection and localization using a two-stage strategy. *arXiv preprint arXiv:1905.00268*, 2019.

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. *arXiv preprint arXiv:1709.06158*, 2017.

Changan Chen, Unnat Jain, Carl Schissler, Sebastià Vicenç Amengual Garí, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip W. Robinson, and Kristen Grauman. SoundSpaces: Audio-visual

navigation in 3d environments. In *Computer Vision – ECCV 2020*, volume 12351 of *Lecture Notes in Computer Science*, pp. 17–36. Springer, 2020. doi: 10.1007/978-3-030-58539-6_2.

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. URL https://openreview.net/forum?id=cR91FAodFMe.

Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W. Robinson, and Kristen Grauman. SoundSpaces 2.0: A simulation platform for visual-acoustic learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8896–8911, 2022. doi: 10.52202/068431-0647. Datasets and Benchmarks Track.

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

Suo Gao, Yueqi Song, Yinghong Cao, Herbert Ho-Ching Iu, Lei Qin, Yushu Zhang, and Jun Mou. Biologically plausible memristive decision-making circuit for adaptive control in industrial autonomous navigation. *IEEE Transactions on Industrial Informatics*, 2026.

Peter Gerstoft, Christoph F Mecklenbräuker, Angeliki Xenaki, and Santosh Nannuru. Multisnapshot sparse bayesian learning for doa. *IEEE Signal Processing Letters*, 23(10):1469–1473, 2016.

Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2616–2625, 2017.

Yicong Hong, Yang Zhou, Ruiyi Zhang, Franck Dernoncourt, Trung Bui, Stephen Gould, and Hao Tan. Learning navigational visual representations with semantic map supervision. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3032–3044. IEEE, 2023.

Maoshen Jia, Yitian Jia, Shang Gao, Jing Wang, and Shusen Wang. Multi-source doa estimation in reverberant environments using potential single-source points enhancement. *Applied Acoustics*, 174:107782, 2021. doi: 10.1016/j.apacoust.2020.107782.

Ziyang Jiang, Yu Chen, Zexu Pan, Xinyuan Qian, Bowen Xing, Ivor W Tsang, Xu-Cheng Yin, and Haizhou Li. Selecttsl: Prompt-guided selective target sound localization in complex scenarios. *arXiv preprint arXiv:2607.02343*, 2026.

Kurt Konolige and Motilal Agrawal. Frameslam: From bundle adjustment to real-time visual mapping. *IEEE Transactions on Robotics*, 24(5):1066–1077, 2008.

Dongheon Lee, Younghoo Kwon, and Jung-Woo Choi. Deepasa: An object-oriented multi-purpose network for auditory scene analysis. *Advances in Neural Information Processing Systems*, 38: 170298–170325, 2026.

Jia Li and Yinfeng Yu. Generalizable audio-visual navigation via binaural difference attention and action transition prediction. In *2026 International Joint Conference on Neural Networks (IJCNN)*, 2026. URL https://arxiv.org/abs/2604.05007.

Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot. Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1997–2012, 2017.

Teng Liu and Yinfeng Yu. Reliability-aware geometric fusion for robust audio-visual navigation. In *2026 International Joint Conference on Neural Networks (IJCNN)*, 2026. URL https://arxiv.org/abs/2604.02391.

Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.

Renana Opochinsky, Mordehay Moradi, and Sharon Gannot. Single-microphone speaker separation and voice activity detection in noisy and reverberant environments. *EURASIP Journal on Audio, Speech, and Music Processing*, 2025(1):18, 2025.

Haotian Pan, Shibo Huang, Jian Yang, Jinpeng Mi, Ke Li, Xiong You, Peidong Liang, Jinbo Yang, Yingjie Liu, Jianfeng Zhang, et al. Robot navigation via foundation language models: A review. *ACM Computing Surveys*, 58(11):1–38, 2026a.

Zexu Pan, Shengkui Zhao, Yukun Ma, Haoxu Wang, Yiheng Jiang, Biao Tian, and Bin Ma. Luseel: Language-queried binaural universal sound event extraction and localization. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 22722–22726. IEEE, 2026b.

Michael Risoud, J-N Hanson, Fanny Gauvrit, Christian Renard, P-E Lemesre, N-X Bonne, and Christophe Vincent. Sound source localization. *European annals of otorhinolaryngology, head and neck diseases*, 135(4):259–264, 2018.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346. IEEE, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, Christoph Feichtenhofer, Piotr Dollár, Wei-Ning Hsu, and Ann Lee. SAM Audio: Segment anything in audio. *arXiv preprint arXiv:2512.18099*, 2025a. URL https://arxiv.org/abs/2512.18099.

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, 2025b. doi: 10.1609/aaai.v39i14.33608.

Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, and Yuki Mitsufuji. Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 915–919. IEEE, 2021.

Hongcheng Wang, Yuxuan Wang, Fangwei Zhong, Mingdong Wu, Jianwei Zhang, Yizhou Wang, and Hao Dong. Learning semantic-agnostic and spatial-aware representation for generalizable visual-audio navigation. *IEEE Robotics and Automation Letters*, 8(6):3900–3907, 2023.

Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019.

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.

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.

Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13032–13056, 2025.

Jinzheng Zhao, Xinyuan Qian, Yong Xu, Haohe Liu, Yin Cao, Davide Berghi, and Wenwu Wang. Text-queried target sound event localization. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pp. 261–265. IEEE, 2024.

Shilian Zheng, Zhuang Yang, Weiguo Shen, Luxin Zhang, Jiawei Zhu, Zhijin Zhao, and Xiaoniu Yang. Deep learning-based doa estimation. *IEEE Transactions on Cognitive Communications and Networking*, 10(3):819–835, 2024.

Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3357–3364. ieee, 2017.

A TARGETDOA: IMPLEMENTATION AND TRAINING DETAILS

A.1 FOA SIGNAL PROCESSING AND ARCHITECTURE

Coordinates and acoustic references. Each observation contains one second of audio at 16 kHz. The native world-frame channels are $[W, Y_{\text{up}}, Z, X]$. Receiver heading is $\psi = \operatorname{atan2}(f_z, f_x)$, defined by the horizontal forward vector. The spatial branch receives $[W, R, U, F]$, with

$$R = -\sin\psi\, X + \cos\psi\, Z, \qquad F = \cos\psi\, X + \sin\psi\, Z, \qquad U = Y_{\text{up}}. \tag{7}$$

Azimuth is zero forward and positive to the right. For each target-only component, the reference uses the window-averaged active-intensity vector $\mathbf{I}_j = [\langle W_j X_j\rangle, \langle W_j Y_j\rangle, \langle W_j Z_j\rangle]^\top$ and subtracts ψ from its horizontal azimuth. Averaging precedes angle extraction. This is a net acoustic arrival direction, not a geometric bearing or a resolved propagation path.

Basic validity requires energy $E_j = \langle W_j^2\rangle \geq 10^{-10}$, $\|\mathbf{I}_j\|_2 \geq 10^{-9}$, and horizontal intensity greater than $\max(10^{-9}, 10^{-6}\|\mathbf{I}_j\|_2)$ for all three source components. Both queried targets additionally require confidence $\min(1, \|\mathbf{I}_j\|_2/E_j) \geq 0.2$, component-energy SIR $10\log_{10}(E_j/\sum_{i\neq j} E_i) \geq -20\,\text{dB}$, and at least 45° circular separation between their references. The SIR denominator excludes cross terms, and confidence is a proxy rather than a calibrated probability. Filtering does not use model predictions or agreement with geometric bearings. Target-only components provide supervision, validity checks, and offline evaluation references, never predictor inputs.

Spatial features and encoding. All four rotated spatial channels share one RMS gain. Its value is $0.1/\max(\text{RMS}(x), 10^{-12})$, with RMS computed over channels and samples. The frozen audio encoder instead uses the original mixture W. The spatial STFT uses a periodic 400-sample Hann window, 160-sample hop, 1024-point FFT, centering, and reflect padding, producing 513 frequency bins and 101 frames. We use 64 non-area-normalized triangular mel filters spanning 50–8000 Hz.

Mel-weighted energy and intensity are formed from $|W|^2$ and $\text{Re}(\overline{W}V^a)$ for $V^a \in \{R, U, F\}$. With numerical floor $\epsilon = 10^{-8}$, intensity magnitude is $n = \sqrt{\sum_a (I^a)^2 + \epsilon}$ and normalized components are $v^a = \text{clip}(I^a/(n+\epsilon), -1, 1)$. The coherence proxy is $q = \frac{1}{2}\,\text{clip}(n/(E+\epsilon), 0, 2)$; diffuseness is $1-q$, and log energy is centered over the time–frequency grid. A reliability head receives these features and clipped horizontal intensity magnitude. Two 3×3, width-32 convolutions with eight-group GroupNorm and GELU, followed by a sigmoid output, produce reliability r without separate supervision. Its evidence weight is $e = r\sqrt{\text{clip}(E/\max(\overline{E}, \epsilon), 0, 10)}$. Here $\bar{E}$ is the mean energy over the time–frequency grid, and L denotes centered log energy.

Two 3×3 convolutions encode the seven channels $[L, v^R, v^U, v^F, q, 1-q, r]$ into width 64, using eight-group GroupNorm and GELU. Frequency pooling adds $\log\max(e, \epsilon)$ to a learned scalar score before softmax. The pooled sequence is projected to width 96 and processed by three kernel-three depthwise temporal residual blocks with dilations 1, 2, and 4 and dropout 0.1, yielding 101 spatial tokens. Frequency pooling is query-independent; target conditioning follows it.

Semantic fusion and prediction. The frozen SAM Audio base encoders provide DAC-VAE features from W resampled to 48 kHz and T5-base features from `the sound of {class}`, preserving underscores in class names. DAC returns 25 continuous pre-quantization mean tokens of width 128; T5 tokens have width 768, with padding masked. No waveform separation model is executed. Linear projections map both sequences to width 96. Text tokens first query audio tokens, then spatial tokens query the resulting semantic context. Both cross-attention blocks use four heads, dropout 0.1, residual connections, and LayerNorm; the second masks padded text positions.

Temporal pooling adds log summed-frequency evidence to a learned score of the conditioned spatial tokens. A $96 \to 253 \to 24$ GELU MLP and softmax produce the posterior, with 280,528 trainable parameters excluding frozen encoders. Training combines equal-weight cross-entropy against normalized $\kappa = 8$ acoustic soft labels and a directional cosine loss. Point directions use the first circular moment rather than bin argmax; its normalization floor is 10^{-6}. No confidence weighting or additional pairwise loss is applied, and near-zero resultants do not invoke a geometric fallback.

Table 5: Standalone FOA data. Each root has one natural and three controlled views, with two target queries per view.

Subset	Scenes	Roots	Views	Queries
Training	59	512	2048	4096
Exposed development	10	32	128	256
Reserved development	Same 10	96	384	768
Canonical final	15	240	960	1920

A.2 FOA Data, Training, and Controls

Paired mixtures and gain views. Each independent mixture root contains two distinct target classes and one background source, with class-to-source-slot assignment randomized. Both queries share the identical waveform, pose, layout, and time window. Natural views disable target-level protection, distance compensation, and background compression. Collection uses fresh native FOA RIRs with temporal coherence disabled and the RIR cache invalidated; exported components are checked against dry-clip convolution and their sum against the mixture. Direct, indirect, diffraction, and transmission paths are enabled, with 500 direct rays, 1024 indirect rays of maximum depth 200, 128 source rays of depth ten, and maximum diffraction order ten. Global volume is one and material mapping is disabled. Repeated native rendering is not assumed to be bitwise deterministic.

Each accepted root also supplies three controlled views without rerendering. For level difference $l \in \{-6, 0, 6\}$ dB and $E_\star = \sqrt{E_0 E_1}$, the desired received W energies are

$$(\widetilde{E}_0, \widetilde{E}_1, \widetilde{E}_b) = E_\star(10^{l/20}, 10^{-l/20}, 10^{-6/10}). \tag{8}$$

Components are scaled by $\sqrt{\widetilde{E}_j / E_j}$, followed if necessary by the common peak gain $\min(1, 0.99/\text{peak})$. These gains preserve reference directions; all views of a root remain in one partition and are not treated as independent observations.

Partitions and scope. Table 5 summarizes the data; final scenes contain 16 roots each. Development uses ten canonical validation scenes, excluding `8194nk5LbLH`. The reference filters define a selected paired-mixture population rather than the full deployment distribution. The target catalogue contains one original recording per class and its slices; background audio contains ten one-second assets. Recordings are shared across partitions: all 128 distinct final target WAVs occur among the 142 training target WAVs, and all ten final background WAVs occur in training. Evaluation therefore concerns closed-set target conditioning under new mixtures and propagation conditions, not unseen recordings or open-vocabulary generalization. Final scenes are disjoint from this head-development data, but had historical exposure elsewhere in the project.

Training and augmentation. All main controls are trained from scratch with seeds 0, 1, and 2 for 40 epochs using AdamW, constant learning rate 3×10^{-4}, weight decay 0.01, batch size 24, and gradient clipping at five. Each epoch visits all 4096 query rows, including the incomplete final batch, for 6840 updates in total. Natural views contribute one quarter of the training weight; controlled views contribute three quarters. One seeded virtual-heading angle δ is shared by both queries and all views of each root within an epoch. It maps the horizontal channels to $R' = \cos\delta\, R - \sin\delta\, F$ and $F' = \sin\delta\, R + \cos\delta\, F$, while subtracting δ from the reference angle. The original W, vertical channel, and cached semantic tokens are unchanged. Evaluation disables random augmentation. Checkpoints are fixed at the final update rather than selected by development performance.

Control definitions and aggregation. The prompt-free control replaces the query with one learned 768-dimensional token while retaining semantic attention. Geometric supervision changes only the reference direction, retaining the same samples, filters, augmentation, and budget. These three arms' nine final checkpoints are fixed before reserved/final evaluation. The supplementary W-only control preserves the query, original DAC input, and trainable architecture, but replaces the spatial input by $[W, 0, 0, 0]$ before RMS normalization and recomputes its features and evidence weights. It uses the same three seeds and update budget. Added after the original results were examined, it is a post-hoc control on the previously used sets, with checkpoints fixed before its evaluation.

It tests directional-input utility, not encoder superiority; its input is invariant to virtual heading while the supervised angle rotates.

Angular errors are shortest circular distances from the posterior-mean direction; Accuracy@30° uses an inclusive acoustic-error threshold. All variants' cross-entropies use the same normalized acoustic soft labels. Queries are averaged within views, and controlled views within roots; root scores are averaged within scenes and then equally across scenes and training seeds. Natural views provide the primary evaluation. No posterior-confidence threshold selects evaluation samples.

A.3 BINAURAL TARGETDOA

Inputs and folded reference. The binaural estimator receives native left–right audio at 16 kHz. The frozen DAC encodes the original left ear and T5 the target prompt; unlike FOA, its spatial branch does not apply common-RMS normalization. Nonfinite waveforms or raw common RMS at most 10^{-7} are invalid. Twelve bin centers $\vartheta_j = -\pi/2 + (j + \frac{1}{2})\pi/12$ cover a folded half-plane at 15° spacing, preserving front–back ambiguity. Training uses a same-frame target-only FOA component as a teacher, with energy/intensity floors 10^{-10}/10^{-9} and confidence at least 0.2. For its world-frame net arrival angle θ^*_{w},

$$\theta^*_{\mathrm{b}} = \operatorname{atan2}\left(\sin(\theta^*_{\mathrm{w}} - \psi), |\cos(\theta^*_{\mathrm{w}} - \psi)|\right). \tag{9}$$

This is an acoustic, not geometric or GCC-PHAT, reference. The corpus contains one primary target query per mixture, unlike the FOA two-query protocol. The extra FOA component is not required at inference.

Features, head, and objective. A 400-sample Hann window, 160-sample hop, 512-point FFT, and 64 mel bands provide normalized real/imaginary interaural cross-spectrum components, right-minus-left log-power ILD, coherence, normalized side power, and centered log power. Reliability has hidden width 32; evidence-weighted frequency pooling, width-96 temporal encoding, two four-head semantic attention blocks, and temporal pooling follow the same structure as the FOA branch, with dropout 0.1 and masking of padded text tokens.

A query-independent pre-fusion token h_{s} supplies a learned $96 \to 96 \to 12$ spatial base. The conditioned token h supplies a bounded correction:

$$\ell = F_{\mathrm{base}}(h_{\mathrm{s}}) + 2g\tanh(W_{\mathrm{res}}h + b_{\mathrm{res}}), \qquad p = \operatorname{softmax}(\ell). \tag{10}$$

The gate uses conditioned and mean semantic tokens, a learned relevance score, and mean reliability/evidence. The residual starts at zero and the gate's final bias at -1. The loss combines equal-weight acoustic soft-label cross-entropy and directional cosine loss, plus weight-0.5 binary cross-entropy between the posterior resultant and detached prediction correctness (angular error below 25°). The resultant is clipped to $[10^{-4}, 1 - 10^{-4}]$ for this auxiliary loss; this objective does not establish calibrated confidence.

Data and training. Of 4000 recorded mixtures, scene splitting and confidence filtering retain 2634 training samples from 59 scenes and 399 validation samples from ten scenes. The corpus uses a 0-dB dry target/background role setting and a 1-dB received target/background protection rule, unlike the unprotected FOA natural views. The available recipe specifies AdamW at 3×10^{-4}, seed 0, batch size 24, clipping at five, and at most 20 epochs, with lowest-validation-median-error selection and patience five. Training swaps ears with probability 0.5, also changing the DAC input and negating the folded reference. The retained history has 15 completed epochs; the selected checkpoint is epoch index nine, after 1100 updates, with 289,662 trainable parameters excluding frozen encoders. The fixed500 scenes participate in frontend selection and are not untouched frontend scene holdouts. Retained artifacts lack a complete optimizer/launch snapshot; defaults not separately logged remain recipe-level specifications.

Table 6: DREAM architecture. Navigation optimization settings are reported separately in Table 7.

Setting	Value
Sectors / history capacity	24 / 16
Token width / attention heads	128 / 4
Depth input / patch grid	128×128 / 16×16
Direction-convolution widths / kernels	$4 \to 64 \to 128$ / 5, 3
Temporal attention blocks / FFN width	2 / 256
Router widths	$131 \to 64 \to 2$
Residual-head widths	$128 \to 202 \to 128 \to 1$
Residual bound / temperature	2 / 1
Prior age decay	0.12

A.4 Posterior Interface to DREAM

FOA outputs a 24-bin full-circle posterior. The binaural adapter divides each folded probability equally between front and rear counterparts:

$$\begin{aligned} p_k^{24} &= \tfrac{1}{2} p_{m(k)}^{12}, \\ m &= (5,4,3,2,1,0,0,1,2,3,4,5, \\ &\quad 6,7,8,9,10,11,11,10,9,8,7,6). \end{aligned} \tag{11}$$

Indices are zero-based. Nonfinite or uniform folded outputs yield a uniform full-circle fallback. This interface preserves ambiguity and passes no separate 96-dimensional frontend token or resultant to DREAM.

B Additional Details of DREAM

DREAM maps the acoustic posterior to a 24-sector navigation preference, not a source coordinate, traversability guarantee, or four-action policy. Its principal architecture settings are in Table 6; optimization is consolidated in Appendix C.2.

B.1 Evidence, Direction Tokens, and Depth Context

The observation adapter normalizes finite log-posteriors and substitutes the uniform distribution u for nonfinite inputs or maximum deviation from u at most 10^{-6}. DREAM treats current evidence as available only when that deviation exceeds $\delta = 2 \times 10^{-6}$. Probability normalization uses $\epsilon = 10^{-8}$ and returns u if total mass is insufficient. External confidence and resultant values are not additional trainable inputs.

The direction encoder receives probability, centered log probability, and two sector-relative projections of the unnormalized circular moment. Two circular convolutions preserve azimuth adjacency. A private depth encoder receives log depth, a broadcast frame-validity flag, two image-coordinate channels, and three calibrated unit-camera-ray channels at 128×128 resolution. Three residual downsampling stages ($7 \to 32 \to 64 \to 96$) and a width-128 projection produce 256 patches. Direction tokens query all patches with four-head cross-attention, without a hard sector-to-patch mask. The flattened camera-to-body transform is embedded and added to depth keys and values. History stores a projected global mean depth token plus this extrinsic embedding. Unavailable depth instead uses normalized direction tokens and a zero global depth context.

B.2 Causal History and Directional Adaptation

Each environment stores up to 16 preceding active states $(q_s, d_s^{\mathrm{g}}, \xi_s, v_s, b_s)$: navigation preference, global depth context, planar pose, directional validity, and depth validity. History stores predictions q_s, not raw arrival posteriors. Age is buffer position, with the newest preceding entry at age zero. For relative yaw $\Delta\psi_{s,t} = \mathrm{wrap}(\psi_s - \psi_t)$ and $\zeta = k - \Delta\psi_{s,t}/(2\pi/K)$, let $j = \lfloor \zeta \rfloor$ and $\eta = \zeta - j$.

Fractional alignment is

$$\widetilde{q}_{s,t}(k) = (1-\eta)q_s(j \bmod K) + \eta q_s((j+1) \bmod K). \tag{12}$$

Only directionally valid entries contribute mass to the age-weighted prior, using decay 0.12; an empty prior is uniform. Relative translation is expressed in the current body frame and embedded with sine/cosine yaw, log displacement magnitude, and sine/cosine displacement bearing. Bearing features are zero at zero displacement. Translation conditions the learned context but does not geometrically reproject historical directions or directly attenuate their prior weights.

Historical tokens add global depth, aligned probability and centered log probability, pose, and learned/normalized age embeddings. Two pre-normalized cross-attention blocks let current tokens query at most 384 historical sector tokens, with learned circular relative-angle biases and width-256 SiLU feed-forward layers. Entries with valid direction or depth provide keys, allowing depth-only history to contribute context without directional prior mass. All-invalid history gives zero attended contribution while preserving the current-token feed-forward path. The current probability and centered log probability are embedded and re-injected before gate and residual prediction.

Routing and correction. The router receives the mean contextual token, normalized entropy-derived evidence, a history-availability flag, and the fraction of directionally valid slots. Its scalar sigmoid outputs are correction and routing gates. Current-only evidence fixes routing to p_t; history-only evidence fixes it to m_t, while either case retains learned correction. If neither is available, correction is disabled and $q_t = u$, even with depth. The entropy-derived statistic is zeroed within 4×10^{-6} of u. For the shared sector head f, the bounded zero-sum residual is

$$\begin{aligned} z_t(k) &= \tanh f(\mathbf{H}_t^{\mathrm{nav}}(k)), \\ c_t(k) &= z_t(k) - K^{-1}\sum_j z_t(j), \qquad r_t(k) = \frac{2c_t(k)}{1+\max_j |c_t(j)|}. \end{aligned} \tag{13}$$

It is applied to the base log distribution with temperature one, as in the main-text directional-reweighting equations. The residual output layer is zero-initialized; correction/routing output biases are $-1/4$, with weight standard deviations $10^{-3}/10^{-4}$. Thus a newly initialized model follows its epsilon-floored base prior; continued checkpoints are not reinitialized.

State and output contract. Prediction reads only preceding states, then appends the current tuple. Buffers reset at episode boundaries, and padding does not write to memory. Stored directional validity is the logical OR of current and historical availability, so history-supported evidence can persist through silence. The 16-slot bound limits explicit states, not the age of all recursively propagated information. RouteFuse receives only q_t and its validity flag, not raw arrival distributions, gates, or separate history tensors.

B.3 Inference-Time Interventions

All variants use one frozen navigation checkpoint and execute their own closed-loop trajectories on identical episodes, without retraining. *Residual-off* sets $g_t = 0$ while retaining routing, prior, and history updates; its output is the normalized epsilon-floored base, not generally p_t. *No-history* replaces DREAM history reads by an empty buffer and sets the history count to zero, retaining current geometry, correction, and the downstream recurrent policy. *TargetDOA-direct* passes $q_t = p_t$ with current-acoustic validity, bypassing all learned routing, correction, and DREAM history effects on the output.

C Fusion and Policy Optimization Details

C.1 RGB-D Projection and Preference-Conditioned Fusion

RouteFuse receives registered 128×128 RGB-D, camera calibration, q_t, and observable self-state. The polar grid has 24 sectors and five radial bins with edges $(0.5, 1.5, 3.0, 5.0, 7.5, 10.5)$ m and centers $(1.0, 2.25, 4.0, 6.25, 9.0)$ m. These bins organize observed geometry, not estimated acoustic range. The RGB residual encoder uses widths 32, 64, and 128 with one, two, and three blocks,

producing $16 \times 16 \times 96$ features from inputs in $[-1, 1]$. The separate navigation-depth encoder uses the seven-channel recipe described for DREAM and produces 256×128 features followed by a linear projection. Its parameters are independent of DREAM's depth encoder.

Support-weighted projection. Each feature patch uses mean positive finite camera-axis depth and its valid pixel fraction. Patches are back-projected through camera intrinsics and transformed to body coordinates. Soft splatting distributes each valid patch to at most two neighboring angular sectors and radial centers, weighted by valid-pixel fraction and interpolation weights. Features are normalized by accumulated support with floor 10^{-8}; support is not binary occupancy. RGB and depth share projection coordinates, with radial, modality, and capped-support embeddings. RGB without accepted radial depth is retained as angular-only camera-ray evidence rather than assigned an artificial range. Projection geometry may be cached detached; feature aggregation remains differentiable.

Directional conditioning. Each cell combines depth, RGB, capped support, and radial index through a $230 \to 128$ projection, SiLU, and LayerNorm. The preference token embeds q_t and sector-relative circular-moment features into width 96, adds relative-angle and validity embeddings, and is projected to width 128. A shared two-layer Transformer processes one preference token and five radial tokens per sector, using four heads, feed-forward width 256, GELU, and zero dropout. The preference output summarizes the sector and receives angular-only RGB evidence. Unsupported radial cells participate internally but retain their base features after attention and context injection. Two residual circular-convolution blocks exchange neighboring-sector context, which is fused with supported radial features.

The resulting cells G_{kr} are modulated by the fixed preference gain

$$\begin{aligned} &\ell_k = \log \max(q_t(k), \epsilon_{\text{tiny}}), \\ &g_k = 1 + 0.5 \tanh\left(\ell_k - K^{-1} \sum_j \ell_j\right), \qquad \widehat{G}_{kr} = g_k G_{kr} + W_\theta [\sin\theta_k, \cos\theta_k]^\top. \end{aligned} \tag{14}$$

Here ϵ_{tiny} is the smallest positive normal value of the probability dtype. Gains lie in $[0.5, 1.5]$ and equal one for uniform q_t; nonuniform preferences need not have unit mean gain.

Policy representation. Two 3×3 convolutional blocks with strides $(2, 1)$ and $(2, 2)$ map the polar tensor through widths 192 and 256 to a 6×3 grid. Flattening and projection produce a 512-dimensional polar feature. A 64-dimensional center token pools pre-projection RGB, depth, and preference features. The self-state branch combines it with modality validity, normalized preference entropy, incremental planar motion, previous action, collision feedback, and elapsed time, mapping 80 inputs to width 128. Polar and self-state features are projected from 640 to 488 dimensions and concatenated with the unmodified 24 probabilities in q_t to form the 512-dimensional policy input. Elapsed time is not goal progress, and the complete architecture is not claimed to be exactly rotation-equivariant.

C.2 Recurrent Optimization and Training Settings

Actor and value branches have independent $516 \to 512$ input projections, one-layer width-512 GRUs, and heads $512 \to 256 \to 4$ and $512 \to 256 \to 1$. They concatenate the previous-action one-hot vector with the fused input; head hidden activations are $\tanh$. Episode boundaries reset recurrent states, and padding does not advance them. The actor updates DREAM, RouteFuse, and trainable RGB-D encoders. The value branch receives detached fused features and updates only its own projection, GRU, and head. The acoustic frontend stays frozen.

Replay and objectives. Each custom rollout contains 32 steps from 64 environments. Episode-local replay uses up to 16 loss-bearing steps and up to 16 preceding context steps, starting from detached entry states. Context steps are loss-masked, not run without gradients: no detach occurs at the context/learning boundary. DREAM and RGB-D features are recomputed causally, while rollout action probabilities, values, and preferences remain fixed PPO references. Active episodes are replayed under updated parameters to refresh recurrent state before further collection.

Table 7: Custom navigation training. Learning rates are constant within each listed stage; shared rollout and optimization settings are specified in text.

Setting	FOA adaptation	Binaural training
Initialization	Navigation checkpoint	From scratch
DREAM learning rate	2.5×10^{-5}	10^{-4}
Other actor/value learning rate	6.25×10^{-5}	2.5×10^{-4}
Action-entropy coefficient	0.01	$0.05 \rightarrow 0.01$
Training seed	11	11

The actor uses the main-text objective in equation 6, with PPO clipping 0.2, action entropy, and weight-0.01 preference KL. Loss averages exclude padding and context prefixes, but do not additionally exclude steps with invalid directional evidence. Preference KL uses float32 and machine-epsilon log flooring; it is not an action-policy KL or geometric supervision term. GAE uses discount 0.99 and parameter 0.95. Continuing episodes bootstrap at rollout boundaries; termination or truncation sets continuation to zero. Advantages are normalized once over all valid rollout transitions before sequence minibatches are formed, without minibatch-wise renormalization. Return targets use the unnormalized GAE advantages plus the rollout-time value estimates. The value loss is one half of the mean maximum of unclipped and clipped squared return errors, with value clipping 0.2. If running mean approximate action-policy KL exceeds 0.02, remaining actor updates for that rollout are skipped, while value updates continue. No geometric-bearing, gate-sparsity, collision-prediction, or modality-classification auxiliary loss is added.

Schedules and checkpoint provenance. The actor uses AdamW with separate DREAM and other-module groups; the value branch uses Adam. DREAM weight decay is 10^{-4}, other groups use zero, optimizer epsilon is 10^{-5}, and actor/value gradient norms are clipped separately to 0.5. Two PPO epochs use global minibatches of 64 loss-bearing steps, sharded over two GPUs without changing that global batch size. Table 7 lists the custom-task schedules. Binaural entropy decreases linearly from 0.05 to 0.01 over 1,048,576 aggregate transitions, then remains at 0.01; FOA adaptation uses fixed 0.01.

The reported FOA model continues an existing navigation checkpoint. Its frozen frontend is replaced after 22,528 adaptation steps while navigation weights, optimizer states, and counters are retained; episodes and causal histories restart. The reported step-262,144 checkpoint therefore contains 239,616 steps after that replacement. These counts describe adaptation, not total training from scratch.

Custom reward. For geodesic distance d_t, the custom navigation reward is

$$r_t = (d_t - d_{t+1}) - 0.01 + 10\mathbf{1}_t^{\text{successful stop}} - \mathbf{1}_t^{\text{timeout}} - 0.05\mathbf{1}_t^{\text{spin}}. \tag{15}$$

Progress is not clipped. A spin requires the current action to be a turn, at least eight turns among retained post-action records with ticks in $[t-10, t]$, and less than 0.1 m net displacement. Timeout is charged when the wrapper reaches 500 steps without an already reported terminal state. There is no additional failed-stop, collision, or wrong-source-stop penalty. Goal distances and success labels are used for rewards and metrics, not policy observations.

D Custom Acoustic Environment: Implementation Details

D.1 Scenes, Sources, and Navigation Protocol

Navigation training uses 59 Matterport3D scenes. The fixed500 development suite contains 50 episodes from each of ten disjoint navigation-evaluation scenes and is used repeatedly for development and checkpoint selection, not as an untouched final test. Each episode contains one stationary target and two to seven stationary background emitters. Background recordings and categories may repeat; emitter counts do not guarantee simultaneous audibility. The target is associated with a target object and corresponding audio/visual prompts. Sound classes are listed in Table 8.

Table 8: Audio-event classes in the custom environment.

Pool	Classes
Target (14)	Bird, dog barking, drum, footstep, guitar, human applause, keyboard typing, man coughing, man talking, piano, telephone ringing, toilet, water tap, woman reading.
Background (10)	Air pump, beeps, big engine, chiller, copy machine, engine, fan, leak, metal door opening, whirr.

Table 9: Custom-navigation renderer settings. Standalone FOA collection instead disables temporal coherence and forces fresh RIRs, as specified in Appendix A.2.

Parameter	Value
Sampling rate / observation length	16 kHz / 1 s
Direct / indirect propagation	Enabled / enabled
Diffraction / transmission	Enabled / enabled
Direct rays	500
Indirect rays / maximum depth	1024 / 200
Source rays / maximum depth	128 / 10
Maximum diffraction order / RIR length	10 / 4 s
Frequency bands / direct–indirect SH orders	4 / 3–1
Temporal coherence	Enabled
Material mapping / mesh simplification	Disabled / disabled
Global volume / unit scale / sensor threads	1 / 1 / 1

Training samples initial geodesic-distance strata $[6, 10)$, $[10, 20)$, and $[20, 50)$ m with probabilities 0.3, 0.4, and 0.3. Minimum oracle-action counts are 8, 15, and 30, with an upper bound of 250 and a minimum geodesic/Euclidean ratio of 1.1 throughout. The development suite contains 150, 200, and 150 episodes in these strata. Forward steps are 1 m, turns are $90°$, and episodes allow 500 actions. Success requires STOP within the configured 1-m geodesic neighborhood of the target, including simulator tolerance. The extra listening action is disabled, but turning advances the acoustic observation.

D.2 Playback and Acoustic Rendering

Source clips are mono, resampled to 16 kHz, and padded or truncated to 16,000 samples. Target clips are ordered by numeric suffix and traversed cyclically. Initialization samples a starting clip and circular within-clip offset; fixed evaluation episodes restore recorded playback state. The offset stays fixed while the target sequence advances. Each background repeats its selected phase-shifted one-second clip. Playback advances with observations rather than wall-clock time, without additional onset, finite-duration, or intermittency gates. Intrinsic silence is retained. This clip-wise process does not guarantee original waveform continuity or amplitude envelopes across boundaries.

Every dry window, including silent samples, is normalized to RMS 0.1 unless its RMS is below 10^{-8}, in which case it is unchanged. Targets and backgrounds have equal dry role gains, not necessarily equal received powers. Background compression and extra distance-based gain compensation are disabled.

Rendering and convolution. Sources are rendered separately at the current receiver pose, with a 1.5-m vertical offset above stored source base positions. Receiver layouts are native binaural LR and four-channel FOA with ACN ordering and N3D normalization; binaural audio is not a downmix of FOA. When both layouts are active, they share sources, windows, and phases but use their respective RIRs. FOA navigation and fixed evaluation also render binaural components for gain control. Binaural-only training omits unused FOA rendering. Online rendering permits reuse of cached source–receiver configurations, using the settings in Table 9.

Linear FFT convolution retains only the first 16,000 samples of each source/window, with zero prehistory and no carried-over waveform tails. Thus the 4-s RIR cap does not imply 4-s waveform memory: delays of at least 1 s do not directly contribute to the retained window. This models within-window multipath, not uninterrupted reverberation across observations. Renderer temporal coherence does not change this truncation, and RIR/ray limits are not measurements of RT_{60}.

D.3 Target Audibility and Mixing

Let $P_{T,t}$ be target binaural power averaged over both channels and the one-second window, and let $P_{B,t}$ be the sum of individual background component powers. For a margin $\tau = 1\,\mathrm{dB}$, backgrounds share the amplitude multiplier

$$a_t = \begin{cases} \min\left(1, \sqrt{P_{T,t}/(10^{\tau/10} P_{B,t})}\right), & P_{T,t} > 0,\ P_{B,t} > 0, \\ 1, & \text{otherwise.} \end{cases} \tag{16}$$

The same multiplier is applied to background components in both FOA and binaural mixtures; it attenuates backgrounds rather than boosting the target. The rule guarantees the configured binaural component-power margin only when both powers are positive. Because the denominator sums individual powers rather than the summed-waveform power, it is neither a guaranteed coherent-mixture SIR nor a guaranteed FOA SIR. This target-dependent protection is part of the evaluation conditions. Each receiver then applies a single common peak gain across samples and channels, preserving its source-component ratios without sample-wise clipping or independent channel normalization.