Listen to Yield: Socially-Aware Audio-Visual Navigation with Non-Line-of-Sight Anticipation

Author Names Omitted for Anonymous Review.

***Abstract*— A service robot often hears social activity before seeing it, such as footsteps behind a wall or a conversation across a doorway. Vision-based social navigation reasons about people only after they enter view, while audio-visual navigation uses sound primarily as a target or interference. Neither turns non-target human sounds into constraints that shape the path. We introduce socially-aware audio-visual navigation (SAVNav), a task and training-free framework with two components: socially-aware audio-visual mapping (SAVMap) and the SAVNav policy. SAVMap anchors persistent sound events to tracked visual entities and propagates unexplained social sounds over the scene topology to anticipate non-line-of-sight (NLOS) human emergence. It then projects the anchoring and NLOS-risk beliefs into a unified social cost field. The SAVNav policy uses this field to couple active sensory exploration with socially-compliant planning. We also construct a 1,008-episode evaluation set in Habitat 3.0 and SoundSpaces 2.0 with online multi-source spatial audio and multi-agent human motion. In occluded scenes, SAVNav raises success rate from 14.0% to 37.5% and lowers human collision rate from 25.3% to 14.3% against a vision-based baseline given ground-truth goal coordinates. The same framework also runs on a physical Stretch 3 robot without parameter retuning.**

I. Introduction

Navigating dynamic, human-populated environments remains difficult for embodied agents and service robots. A robot must avoid collisions, respect implicit social norms, and still complete its task. It should keep distance from a conversation, yield to someone approaching from a blind corner, and respond to an acoustic target such as a doorbell or human call. Humans routinely combine visual and acoustic cues in these situations. Current navigation systems keep the modalities apart: audio serves mainly as a beacon toward a goal, while social reasoning remains visual.

Two navigation lines have complementary failure modes (Fig. 1). Vision-based social navigation (SocialNav) [1]–[3] forecasts pedestrians from visual observations. Occlusion-aware variants can infer hidden agents indirectly from interactions among visible people [4], but cannot exploit the hidden person's own footsteps or conversation. They therefore cannot use those cues to place risk at a likely blind-corner emergence location or establish a social boundary around a conversational group before visual observation. Audio-visual navigation (AVNav) [5]–[10] locates sound sources but represents them only as goals or interference. A human sound therefore cannot constrain its path or encode a boundary around an ongoing activity, so the blue path cuts through the conversational group.

We therefore introduce socially-aware audio-visual navigation (SAVNav), built on a physical premise: sound reaches around corners and through doorways where light cannot. A spatial boundary can thus be inferred before the person forming it is seen. SAVNav turns transient sounds into persistent spatial state, allowing the robot to yield before a non-line-of-sight (NLOS) human appears while still approaching an acoustic target. The main contributions are:

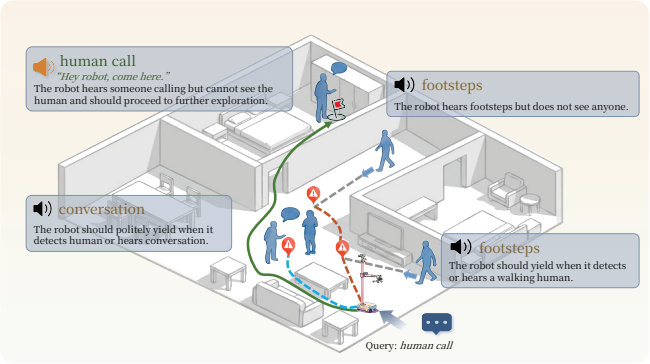


Fig. 1. Navigation to a calling human (flag). Callouts identify audible events, gray dashed lines show human trajectories, and red badges mark a collision or personal-space violation. Vision-based navigation (orange) misses the non-line-of-sight (NLOS) pedestrian at the doorway; audio-visual navigation (blue) cuts through the conversational group. SAVNav (green) anticipates the pedestrian and yields wide of the group.

1) *The SAVNav task:* an embodied navigation setting in which the agent locates an acoustic target while treating non-target sound as a boundary to respect, rather than a goal to reach or noise to suppress. The agent must yield to line-of-sight (LOS) and NLOS humans it may only hear.
2) *SAVMap:* a spatio-temporal representation that binds each persistent sound to a tracked LOS human or static object through audio-visual anchoring. When no LOS entity explains a social sound, topology-aware acoustic anticipation assigns its risk to likely NLOS emergence nodes. Both forms of evidence enter one continuous social cost field.
3) *SAVNav policy:* a training-free planner that trades expected target detection against travel time to select informative viewpoints while the target is unanchored. Once anchored, the target becomes the goal. The social cost field slows and bends the path in both stages.
4) *Dataset and findings:* a 1,008-episode evaluation set couples continuous multi-source spatial audio with multi-agent human motion in Habitat 3.0 and SoundSpaces 2.0. Its metrics isolate NLOS collisions. SAVNav leads all baselines on success and efficiency across three scenarios; in occluded scenes, it reaches

37.5% success versus 14.0% and 14.3% human collisions versus 25.3% for a vision baseline given ground-truth goals. The framework also runs on a physical Stretch 3 robot with the same parameters.

II. Related Work

A. Vision-Based Social Navigation

SocialNav extends collision avoidance to proxemics, path etiquette, and context-dependent social behavior [1]. Recent approaches forecast future pedestrian states and reserve space around them [2] or use vision-language models to score candidate paths [3]. Occlusion-aware crowd navigation can further infer hidden agents from interactions among visible people [4], showing that vision need not be strictly LOS. However, such inference depends on visible proxy actors rather than direct evidence from the hidden person. It cannot exploit that person's footsteps or conversation to place risk at a likely emergence location before visual observation.

B. Audio-Visual Navigation

AVNav uses omnidirectional hearing to overcome a restricted visual field. Early work navigates through unexplored environments toward a continuously emitting, static acoustic target [5]. Later systems relax this assumption by using transient semantic sounds [6], tracking moving sound sources [8], or separating a target from multi-source interference [10]. Other formulations query conversational oracles [7] or use large language models for exploration and layout reasoning [9]. Across these settings, non-target sounds remain alternative goals, interaction cues, or interference. They do not become spatial social boundaries that can redirect the path around a conversation or an NLOS pedestrian.

C. Multi-Modal Spatial Mapping

Contrastive language-audio pretraining (CLAP) [11] and ImageBind [12] provide open-vocabulary cross-modal alignment. They support semantic sound-event localization and detection [13], but these extractors operate on individual observation windows. They do not by themselves turn transient detections into persistent spatial constraints.

Spatial mapping instead aggregates observations into dense fields or sparse topologies. Dynamic and social scene graphs persist people, actions, and relations for downstream reasoning [14], [15], but typically omit localized auditory evidence. Classical acoustic mapping accumulates directional likelihoods onto geometric structures through ray casting [16], but represents sound-emitting structure probabilities rather than dynamic social entities. AVLMaps [17] registers a sound at the agent's location once a decibel threshold is exceeded. This proximity-logging assumption cannot infer a distant source from its bearing. HomeEmergency [18] updates probability-based scene graphs from audio, but localizes events through idealized directionality and predefined human-activity heatmaps. SAVMap instead maintains transient sound events as persistent visual or topological hypotheses and projects both observed and inferred entities into a continuous social cost field for proactive planning.

III. Problem Formulation

The SAVNav task takes place in a continuous 3D indoor environment with static objects $\mathcal{O}$ and human actors $\mathcal{H}$. Exactly one target sound event marks the goal, with label in $\mathcal{L}_{\text{target}} = \{\texttt{doorbell}, \texttt{human call}, \texttt{television}, \texttt{water tap}\}$. Only `human call` is emitted by some $h \in \mathcal{H}$; the other targets come from static objects. Concurrent social sound events have labels in $\mathcal{L}_{\text{social}} = \{\texttt{footsteps}, \texttt{conversation}\}$ and are always emitted by humans. No goal is marked by social sound events; instead, they provide auditory evidence of human activity.

The given static semantic map $\mathcal{M}$ contains a 2D navigable grid $\mathcal{M}_{\text{nav}}$ and the set of static objects $\mathcal{O}$. The grid $\mathcal{M}_{\text{nav}} \subset \mathbb{R}^2$ lists occupiable ground positions. Each object has a 3D center $\mathbf{p}_{\text{obj}}$ and an offline ImageBind [12] embedding $\mathbf{e}_{\text{obj}}$.

Each episode specifies a target-class query $q \in \mathcal{L}_{\text{target}}$. At step t, the observation $O^{(t)} = (I^{(t)}, D^{(t)}, W^{(t)}, \xi^{(t)})$ contains an RGB image $I^{(t)}$, its aligned depth map $D^{(t)}$, a multi-channel waveform $W^{(t)}$, and the localized $\mathrm{SE}(3)$ pose $\xi^{(t)}$; $\mathbf{p}_{\text{robot}}^{(t)}$ is its position. The action space $\mathcal{A} = \mathcal{U} \cup \{\texttt{stop}\}$ combines base commands $\mathbf{u}^{(t)} = (v^{(t)}, \omega^{(t)}) \in \mathcal{U} = [0, v_{\max}] \times [-\omega_{\max}, \omega_{\max}]$ with the arrival command `stop`. Here, $v^{(t)}$ and $\omega^{(t)}$ are the base linear and angular velocities, bounded by $v_{\max}$ and $\omega_{\max}$, respectively. The policy is

$$\pi : \big(O^{(1:t)},\ \mathcal{M},\ q\big) \mapsto a^{(t)} \in \mathcal{A}, \tag{1}$$

where $O^{(1:t)}$ is the observation history. The terminal step t_{end} occurs at budget $T = 500$, human collision, or `stop`. Success requires

$$\big(a^{(t_{\text{end}})} = \texttt{stop}\big) \ \wedge\ \big(\|\mathbf{p}_{\text{robot}}^{(t_{\text{end}})} - \mathbf{p}_{\text{target}}\| \le d_{\text{success}}\big), \tag{2}$$

where the ground-truth target position $\mathbf{p}_{\text{target}}$ is used only for evaluation and $d_{\text{success}} = 1.0\,\text{m}$.

All spatial operations use $\mathbf{x} \in \mathcal{M}_{\text{nav}}$, with 3D quantities projected vertically. We define $\hat{\mathbf{a}} = \mathbf{a}/\|\mathbf{a}\|$, $\angle(\mathbf{a}, \mathbf{b}) = \arccos(\hat{\mathbf{a}}^\top \hat{\mathbf{b}})$ in radians, and $[z]_0^1 = \min(1, \max(0, z))$.

IV. Method

SAVNav maintains spatial social boundaries from transient per-step sounds through three modules (Fig. 2). Multi-modal perception tracks human entities and sound events. SAVMap associates them across modalities and persists the result in a social cost field covering LOS and NLOS humans. The SAVNav policy plans over this field, choosing its goal according to target anchoring. All modules execute synchronously once per control step at 20 Hz.

A. Multi-Modal Perception

1) Visual Perception: YOLO26 [19] detects and segments humans in $(I^{(t)}, D^{(t)})$ in real time. Depth back-projection gives position $\mathbf{p}_{\text{hum}}^{(t)}$ and uncertainty $\sigma_{\text{hum}}^{(t)}$ from squared depth and mask coverage of the detection box. ImageBind [12] encodes the crop as $\mathbf{e}_{\text{hum}}^{(t)}$, and differentiating the smoothed cross-frame track yields velocity $\mathbf{v}_{\text{hum}}^{(t)}$. These attributes form $E_{\text{hum}}^{(t)} = (\mathbf{p}_{\text{hum}}^{(t)}, \mathbf{v}_{\text{hum}}^{(t)}, \sigma_{\text{hum}}^{(t)}, \mathbf{e}_{\text{hum}}^{(t)}) \in \mathcal{E}_{\text{hum}}^{(t)}$.

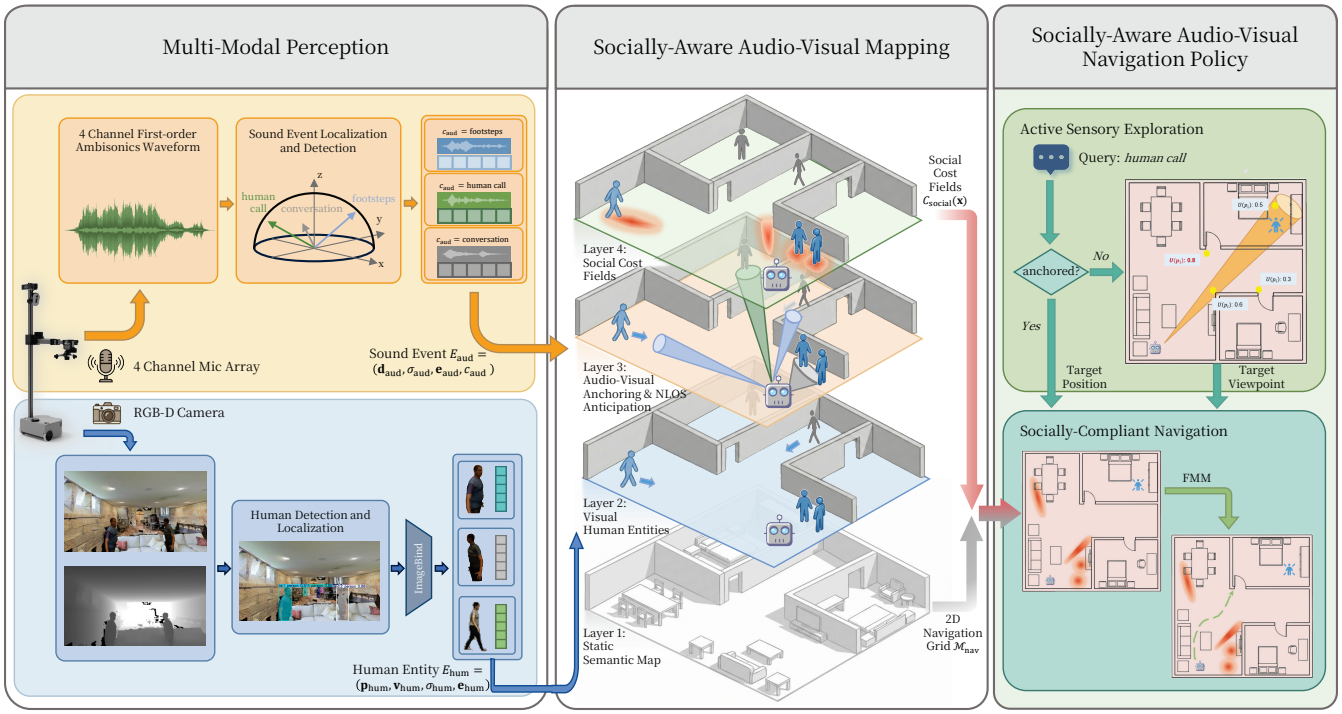


Fig. 2. SAVNav architecture. Perception extracts sound and human entities; socially-aware audio-visual mapping (SAVMap) forms a social cost field for line-of-sight (LOS) humans and topologically inferred non-line-of-sight (NLOS) risks. The policy selects viewpoints until the target anchors, then plans a socially-compliant path with the Fast Marching Method (FMM).

2) Acoustic Perception: The continuous first-order Ambisonics (FOA) waveform $W^{(t)}$ enters embed-ACCDOA [13] through a 2.55 s sliding window inferred every step, so adjacent windows overlap heavily. A window may return several detections, each with an egocentric direction-of-arrival (DOA) and CLAP embedding [11]. Its nearest text embedding among $\{q\} \cup \mathcal{L}_{\text{social}}$ gives class $c^{(t)}_{\text{aud}}$; ImageBind encodes that label as $\mathbf{e}^{(t)}_{\text{aud}}$, sharing a semantic space with visual entities. Pose $\xi^{(t)}$ converts the DOA to world direction $\mathbf{d}^{(t)}_{\text{aud}}$, with confidence-scaled angular uncertainty $\sigma^{(t)}_{\text{aud}}$. Adjacent detections with similar directions and CLAP embeddings are associated across windows. This produces persistent events $E^{(t)}_{\text{aud}} = (\mathbf{d}^{(t)}_{\text{aud}}, \sigma^{(t)}_{\text{aud}}, \mathbf{e}^{(t)}_{\text{aud}}, c^{(t)}_{\text{aud}}) \in \mathcal{E}^{(t)}_{\text{aud}}$.

B. Socially-Aware Audio-Visual Mapping (SAVMap)

SAVMap maps $\mathcal{E}^{(t)}_{\text{aud}}$, $\mathcal{E}^{(t)}_{\text{hum}}$, and $\mathcal{M}$ to $C^{(t)}_{\text{social}}$. It recursively maintains two belief matrices. $\mathbf{B}^{(t)}_{\text{anch}}$ records which visual candidate likely emitted each sound; $\mathbf{B}^{(t)}_{\text{nlos}}$ records where each unanchored social emitter may emerge. Indices k, m, and j denote sounds, candidates, and nodes; $K^{(t)}$ and $M^{(t)}$ are the sound and candidate counts.

1) Audio-Visual Anchoring: Anchoring identifies the observed emitter. Candidate set $\mathcal{E}^{(t)}_{\text{vis}} = \mathcal{E}^{(t)}_{\text{hum}} \cup \mathcal{O}$ represents candidate m by position $\mathbf{p}^{(t)}_{\text{vis},m}$, velocity $\mathbf{v}^{(t)}_{\text{vis},m}$, uncertainty $\sigma^{(t)}_{\text{vis},m}$, and embedding $\mathbf{e}^{(t)}_{\text{vis},m}$. Static objects have zero velocity and uncertainty. Semantic evidence is cosine similarity $s^{(t)}_{km} \in [-1, 1]$ between $\mathbf{e}^{(t)}_{\text{aud},k}$ and $\mathbf{e}^{(t)}_{\text{vis},m}$. Geometric evidence is the DOA alignment

$$g^{(t)}_k(\mathbf{p}) = \exp\left(-\frac{\angle\left(\mathbf{p} - \mathbf{p}^{(t)}_{\text{robot}},\ \mathbf{d}^{(t)}_{\text{aud},k}\right)^2}{2\left(\sigma^{(t)}_{\text{aud},k}\right)^2}\right) \in (0, 1], \quad (3)$$

where $\mathbf{p}$ is a candidate source position; the weight peaks at 1 on sound k's DOA and widens with $\sigma^{(t)}_{\text{aud},k}$.

First-order Markov smoothing accumulates noisy per-step evidence. Element (k, m) of $\mathbf{B}^{(t)}_{\text{anch}} \in [0, 1]^{K^{(t)} \times M^{(t)}}$ is the confidence that candidate m emitted sound k:

$$B^{(t)}_{\text{anch}}(k, m) = \Big[\lambda_{\text{anch}} B^{(t-1)}_{\text{anch}}(k, m) + \eta_{\text{anch}}\, s^{(t)}_{km}\, g^{(t)}_k\left(\mathbf{p}^{(t)}_{\text{vis},m}\right)\Big]^1_0, \quad (4)$$

Here $\lambda_{\text{anch}} = 0.9$ forgets old evidence and $\eta_{\text{anch}} = 0.5$ absorbs the current observation; new rows and columns start at zero.

If $\max_m B^{(t)}_{\text{anch}}(k, m) > \tau_{\text{anch}} = 0.5$, sound k anchors to the arg-max candidate and inherits its position and uncertainty. Otherwise it remains unanchored. The candidate linked to query q becomes target entity $E^{(t)}_{\text{tgt}}$ at $\mathbf{p}^{(t)}_{\text{tgt}}$.

2) Topology-Aware Acoustic Anticipation: An unanchored social sound usually indicates an NLOS emitter. SAVMap assigns its directional evidence to J topological nodes $\mathcal{N}_{\text{topo}} = \{n_1, \ldots, n_J\}$ at positions $\mathbf{p}_{n_j}$. These corners, doorways, and junctions are extracted offline from $\mathcal{M}_{\text{nav}}$. An NLOS human must pass through such a location to enter view, so the nodes turn a bearing into candidate emergence positions. Node count varies by scene, with median 83.

Element (k, j) of $\mathbf{B}_{\text{nlos}}^{(t)} \in [0,1]^{K^{(t)} \times J}$ is the likelihood that sound k's emitter emerges at n_j. Only rows for unanchored social sounds are updated. The update reuses (3), omits semantic similarity, and attenuates distant nodes:

$$B_{\text{nlos}}^{(t)}(k,j) = \Big[\lambda_{\text{nlos}} B_{\text{nlos}}^{(t-1)}(k,j) + \eta_{\text{nlos}}\, g_k^{(t)}(\mathbf{p}_{n_j}) \exp\left(-\frac{\|\mathbf{p}_{n_j} - \mathbf{p}_{\text{robot}}^{(t)}\|^2}{2\sigma_{\text{dist}}^2}\right)\Big]_0^1, \quad (5)$$

Here $\lambda_{\text{nlos}} = 0.9$, $\eta_{\text{nlos}} = 0.3$, and distance scale $\sigma_{\text{dist}} = 4.0\,\text{m}$. A DOA provides bearing but not range, so the distance factor concentrates new evidence on nearer, more urgent nodes. Anchoring or sound cessation clears the row; an anchored event instead contributes through its entity.

For each event, $j_k^* = \arg\max_j B_{\text{nlos}}^{(t)}(k,j)$. Maxima above $\tau_{\text{nlos}} = 0.4$ form the active emergence set

$$\mathcal{N}^{*(t)} = \big\{(k,j) \;\big|\; j = j_k^*,\ B_{\text{nlos}}^{(t)}(k,j) > \tau_{\text{nlos}}\big\}, \quad (6)$$

with at most one node per event.

3) Social Cost Field: The social cost field adds observed-human and active-node components. Both use the same kernel, with observed LOS or inferred NLOS source states:

$$\kappa(\mathbf{x}; \mathbf{p}, \mathbf{v}, a, r_\perp) = a \exp\left(-\frac{\delta_\parallel^2}{2r_\parallel^2} - \frac{\delta_\perp^2}{2r_\perp^2}\right), \quad (7)$$

where $\mathbf{p}$, $\mathbf{v}$, a, and $r_\perp$ are source position, velocity, amplitude, and lateral extent. In the velocity-aligned frame, $\delta_\parallel = (\mathbf{x} - \mathbf{p})^\top \hat{\mathbf{v}}$ and $\delta_\perp = \|(\mathbf{x} - \mathbf{p}) - \delta_\parallel \hat{\mathbf{v}}\|$. Ahead of a moving source, $r_\parallel = r_\perp(1 + k_{\text{v}}\|\mathbf{v}\|)$ with $k_{\text{v}} = 1.5\,\text{s}$; otherwise $r_\parallel = r_\perp$. For $\mathbf{v} = \mathbf{0}$, we set $\delta_\parallel = 0$ and $\delta_\perp = \|\mathbf{x} - \mathbf{p}\|$, giving an isotropic Gaussian. Summing over both source sets gives

$$\begin{aligned} C_{\text{los}}^{(t)}(\mathbf{x}) &= \sum_{E_{\text{vis},m}^{(t)} \in \mathcal{E}_{\text{hum}}^{(t)} \setminus \{E_{\text{tgt}}^{(t)}\}} \kappa\big(\mathbf{x}; \mathbf{p}_{\text{vis},m}^{(t)}, \mathbf{v}_{\text{vis},m}^{(t)}, A, r_{\perp,m}^{(t)}\big), \\ C_{\text{nlos}}^{(t)}(\mathbf{x}) &= \sum_{(k,j) \in \mathcal{N}^{*(t)}} \kappa\big(\mathbf{x}; \mathbf{p}_{n_j}, \mathbf{v}_{*kj}^{(t)}, A\, B_{\text{nlos}}^{(t)}(k,j), r_{\perp,k}^{(t)}\big), \\ C_{\text{social}}^{(t)}(\mathbf{x}) &= \big[C_{\text{los}}^{(t)}(\mathbf{x}) + C_{\text{nlos}}^{(t)}(\mathbf{x})\big]_0^1, \end{aligned} \quad (8)$$

Here $A = 2.0$ is the base amplitude. $C_{\text{los}}^{(t)}$ excludes the target entity, and clamping confines the sum to $[0,1]$.

A human uses its measured state, amplitude A, and $r_{\perp,m}^{(t)} \propto \sigma_{\text{vis},m}^{(t)}$ with lower bound $0.85\,\text{m}$. Together with $A > 1$, this saturates cost $1.0\,\text{m}$ from the body, forming a unit-cost soft wall. Node (k,j) uses $\mathbf{p}_{n_j}$, amplitude $AB_{\text{nlos}}^{(t)}(k,j)$, and $r_{\perp,k}^{(t)} \propto \sigma_{\text{aud},k}^{(t)}$. For `footsteps`, $\mathbf{v}_{*kj}^{(t)} = v_{\text{walk}}(\widehat{\mathbf{p}_{\text{robot}}^{(t)} - \mathbf{p}_{n_j}})$ with $v_{\text{walk}} = 1.2\,\text{m/s}$; for `conversation`, $\mathbf{v}_{*kj}^{(t)} = \mathbf{0}$.

C. Socially-Aware Audio-Visual Navigation Policy (SAVNav policy)

Target anchoring changes only the navigation goal; the social cost field always constrains the path.

1) Active Sensory Exploration: While the target is unanchored, $P_{\text{tgt}}^{(t)}(\mathbf{x}) \in [0,1]$ records its grid belief, initialized by $P_0 = 0.01$. Positive evidence comes from audition; negative evidence comes from observing a cell without the target. Let k_{tgt} index the sound matching query q. Auditory evidence acts on sector $\Omega_{\text{tgt}}^{(t)} = \{\mathbf{x} \in \mathcal{M}_{\text{nav}} \mid \angle(\mathbf{x} - \mathbf{p}_{\text{robot}}^{(t)}, \mathbf{d}_{\text{aud},k_{\text{tgt}}}^{(t)}) \le \zeta \sigma_{\text{aud},k_{\text{tgt}}}^{(t)}\}$ with $\zeta = 0.5$. Visual absence acts on $\mathcal{V}^{(t)}$, the current camera view within $\mathcal{S}(\mathbf{p}_{\text{robot}}^{(t)})$. Here $\mathcal{S}(\mathbf{p})$ denotes navigable cells visible from $\mathbf{p}$ within $R_{\max} = 5.0\,\text{m}$. The update is

$$P_{\text{tgt}}^{(t)}(\mathbf{x}) = \begin{cases} \big[P_{\text{tgt}}^{(t-1)}(\mathbf{x}) + \eta_+\, g_{k_{\text{tgt}}}^{(t)}(\mathbf{x})\big]_0^1, & \mathbf{x} \in \Omega_{\text{tgt}}^{(t)}, \\ 0, & \mathbf{x} \in \mathcal{V}^{(t)} \setminus \Omega_{\text{tgt}}^{(t)}, \\ \lambda_{\text{decay}} P_{\text{tgt}}^{(t-1)}(\mathbf{x}), & \text{otherwise}, \end{cases} \quad (9)$$

where $\eta_+ = 0.2$ is the positive increment and $\lambda_{\text{decay}} = 0.983$ the per-step decay (half-life $2.0\,\text{s}$). Visual absence clears a cell; the set difference prevents conflicting updates. Without target audio, $\Omega_{\text{tgt}}^{(t)} = \varnothing$.

Candidate set $\mathcal{P}^{(t)} = \{\mathbf{p}_i^{(t)}\}_{i=1}^{N_{\text{v}}^{(t)}}$ combines topological-node positions, $N_{\text{roi}} = 10$ navigable samples in $\Omega_{\text{tgt}}^{(t)}$, and the belief centroid, so $N_{\text{v}}^{(t)}$ varies. Candidate i has utility

$$U^{(t)}(\mathbf{p}_i^{(t)}) = \sum_{\mathbf{x} \in \mathcal{S}(\mathbf{p}_i^{(t)})} P_{\text{tgt}}^{(t)}(\mathbf{x}) - \beta \frac{L(\mathbf{p}_{\text{robot}}^{(t)}, \mathbf{p}_i^{(t)})}{v_{\max}}, \quad (10)$$

where $L(\cdot,\cdot)$ is geodesic distance and $\beta = 0.4$ weights travel time $L/v_{\max}$. The maximizer $\mathbf{p}^{*(t)}$ becomes the current goal.

2) Socially-Compliant Navigation: The goal $\mathbf{x}_{\text{goal}}^{(t)}$ is anchored target $\mathbf{p}_{\text{tgt}}^{(t)}$ or viewpoint $\mathbf{p}^{*(t)}$. Fast Marching Method (FMM) [20] solves $\|\nabla \mathcal{T}^{(t)}(\mathbf{x})\| F^{(t)}(\mathbf{x}) = 1$ for arrival time $\mathcal{T}^{(t)}$ from that goal. Its speed field decreases linearly with social cost:

$$F^{(t)}(\mathbf{x}) = F_{\min} + (F_{\max} - F_{\min})\big(1 - C_{\text{social}}^{(t)}(\mathbf{x})\big), \quad (11)$$

Here $F_{\max} = 1.1$ is the propagation ceiling without social cost and $F_{\min} = 0.15$ the positive minimum, both dimensionless relative traversabilities. The positive floor keeps high-cost regions reachable, making social cost a soft constraint rather than a hard obstacle. Gradient descent on $\mathcal{T}^{(t)}$ gives the reference path, which the local controller tracks through $\mathbf{u}^{(t)}$. Once the target is anchored and $\|\mathbf{p}_{\text{robot}}^{(t)} - \mathbf{x}_{\text{goal}}^{(t)}\| \le d_{\text{success}}$, the policy emits `stop`.

V. Dataset

A. Simulation Environment

We instantiate the SAVNav task in Habitat 3.0 [21], where a velocity-controlled robot, human motion, and emitted sounds evolve in one loop. SoundSpaces 2.0 [22] renders each source online, so the FOA stream follows dynamic pedestrians rather than replaying precomputed responses.

Following AudioSet [23], we use four target events (`doorbell`, `human call`, `television`, and `water tap`) and two social events (`conversation` and

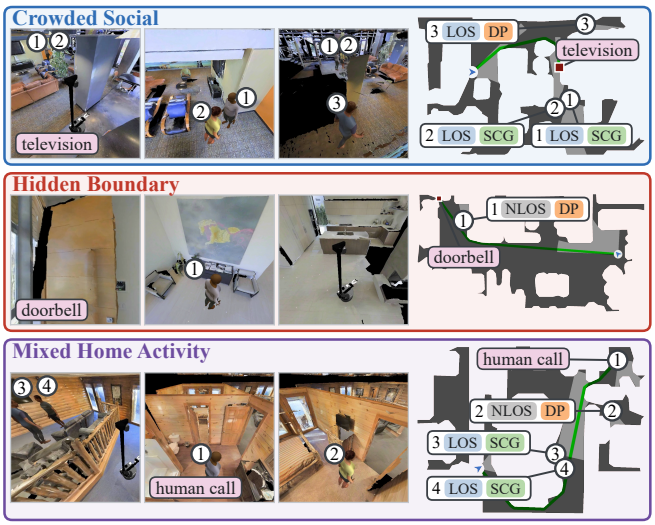


Fig. 3. Representative episodes. Each panel pairs three views with a ground-truth map; the green line is the shortest path from the blue start pose to the pink acoustic target. Badges identify human actors by line-of-sight (LOS) or non-line-of-sight (NLOS) visibility and role: stationary conversational group (SCG) or dynamic pedestrian (DP).

`footsteps`). The 176 recordings come from BeDAViN [10], FSD50K [24], and Freesound[1]. An episode loops up to three concurrent sources. The acoustic target is attached to its static object, except `human call`, which is carried by a stationary actor. Social sounds belong only to humans, and `footsteps` play only while a pedestrian walks. Every source is convolved with its current room response before superposition. The resulting continuous 4-channel FOA stream exposes NLOS risks through audio rather than privileged annotations.

B. Scenarios

Three scenarios separate LOS social routing from NLOS anticipation (Fig. 3). Counts exclude the stationary actor added by a `human call` target.

1) *Crowded Social:* Two to four LOS distractors (mean 3.2) form one stationary conversational group (SCG) of two or three people. They face a shared centroid within $1.25\,\mathrm{m}$ and jointly emit `conversation`; at most one dynamic pedestrian (DP) is added. They occupy main routes, requiring socially-compliant passage.
2) *Hidden Boundary:* One to three NLOS distractors form either one approaching DP or a DP and a two-person SCG. Raycasts enforce initial occlusion. Waypoints make the DP cross the shortest path once the agent is within $3.0\,\mathrm{m}$, so ignoring footsteps cannot avoid the encounter.
3) *Mixed Home Activity:* Three to four distractors (mean 3.7) combine one intersecting NLOS DP with one LOS SCG. This tests whether observed and inferred constraints can be balanced in the same episode.

[1]`https://freesound.org/`

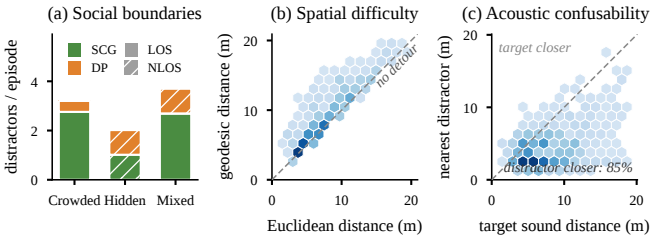


Fig. 4. Statistics for the 1,008 episodes. (a) Scenario composition by stationary conversational group (SCG), dynamic pedestrian (DP), and line-of-sight (LOS) or non-line-of-sight (NLOS) state. (b) Geodesic versus Euclidean start–target distance; points above the diagonal require detours. (c) Target versus nearest-distractor geodesic distance; a distractor is closer in 85% of episodes.

C. Episode Generation

We use Matterport3D [25] environments, restricting each episode to one floor and at most 10 rooms. The 1,008 episodes span 10 scenes and balance the three scenarios (336 each) and four targets (252 each).

Fig. 4 verifies the intended density and visibility. Most episodes lie above the geodesic–Euclidean diagonal, so reaching the target requires routing around the layout rather than following the acoustic bearing. A distractor is geodesically closer than the target in 85% of episodes. Proximity alone therefore cannot separate target and social sounds.

VI. Experiments

A. Experimental Setup

The agent is a Stretch 3 robot[2] with forward-facing RGB-D and a 4-channel Ambisonic array. Method parameters remain fixed across scenarios. Simulation and hardware differ only in $v_{\max}$ (1.7 versus $0.3\,\mathrm{m/s}$); both use $\omega_{\max} = 1.9\,\mathrm{rad/s}$.

The vision-only *Falcon* [2] never uses social sounds as NLOS cues. Two variants separate goal finding from social safety: *Falcon+Goal Oracle* receives ground-truth target coordinates, while *Falcon+Audio Goal* receives our audio anchor. *ENMuS*^{3} [10] follows target audio but ignores human sounds as boundaries. Its released weights assume a fixed BeDAViN vocabulary. We therefore use the 504 episodes with `water tap` or `television`; scenes, geometry, and social boundaries are unchanged. *Reactive SAVNav* uses our perception without inter-frame state ($\lambda_{\mathrm{nlos}} = 0$ in (5), $B^{(t-1)}_{\mathrm{anch}} = 0$ in (4), and no $P^{(t)}_{\mathrm{tgt}}$ in (10)). It matches sounds independently per frame, placing each one $2.0\,\mathrm{m}$ along its DOA.

B. Evaluation Metrics

We measure completion, safety, and social compliance. The metrics are success rate (SR), success weighted by path length (SPL), success weighted by time length (STL), and progress. We also report human collision rate (HCR), NLOS collision rate (NCR), and personal space compliance (PSC):

[2]`https://hello-robot.com/stretch-3-product`

TABLE I

EVALUATION AND ABLATIONS ON 1,008 EPISODES (PERCENT; ↑/↓: HIGHER/LOWER IS BETTER). METRICS ARE SUCCESS (SR), PATH/TIME EFFICIENCY (SPL/STL), PROGRESS, HUMAN/NLOS COLLISION RATES (HCR/NCR), AND PERSONAL-SPACE COMPLIANCE (PSC). BOLD MARKS THE BEST FULL-SET VALUE; DASHES ARE INAPPLICABLE. ENMUS^{3†} USES THE 504-EPISODE SUBSET AND IS EXCLUDED FROM BOLDING.

Method	Crowded Social							Hidden Boundary							Mixed Home Activity						
	SR↑	SPL↑	STL↑	Prog.↑	HCR↓	NCR↓	PSC↑	SR↑	SPL↑	STL↑	Prog.↑	HCR↓	NCR↓	PSC↑	SR↑	SPL↑	STL↑	Prog.↑	HCR↓	NCR↓	PSC↑
Baselines & Ours																					
Falcon+Goal Oracle	8.04	5.81	2.73	34.71	30.95	–	89.82	13.99	11.12	5.41	39.79	25.30	8.33	95.01	10.42	7.84	3.70	32.51	38.99	11.31	87.44
Falcon+Audio Goal	7.74	6.62	3.53	30.27	33.33	–	87.83	10.42	8.17	4.37	34.87	27.08	11.31	94.82	6.55	5.55	2.88	29.87	38.69	11.61	88.50
ENMuS^{3†}	0.00	0.00	0.00	9.93	14.88	–	95.67	0.00	0.00	0.00	18.05	10.12	1.79	98.99	0.00	0.00	0.00	16.62	26.79	4.17	93.84
Reactive SAVNav	23.21	20.93	15.62	42.09	12.80	–	94.19	22.62	21.02	16.30	43.09	**13.39**	5.95	96.66	25.00	22.64	17.31	43.05	15.77	5.95	95.03
SAVNav (Ours)	**37.80**	**34.20**	**25.68**	**49.77**	**12.20**	–	**94.75**	**37.50**	**35.68**	**28.28**	**51.58**	14.29	**5.65**	**97.62**	**39.29**	**36.05**	**28.08**	**51.39**	**13.39**	**5.65**	**96.17**
Ablation Variants																					
w/o Topology-Aware Anticipation	–	–	–	–	–	–	–	34.23	32.48	26.13	51.49	19.05	6.85	95.97	38.39	35.56	27.09	51.29	14.58	6.55	95.51
w/o Active Sensory Exploration	–	–	–	–	–	–	–	36.01	33.85	27.04	50.59	14.58	6.55	96.62	36.01	33.47	26.34	49.48	14.29	6.25	95.83

1) *SR:* episodes stopped within 500 steps and $1.0\,\mathrm{m}$ of the target without human collision.
2) *SPL:* success weighted by shortest-path efficiency.
3) *STL:* mean $S_e T_e^* / \max(T_e^*, T_e)$ over episodes e, where S_e is success and T_e, T_e^* are elapsed and shortest-path steps. It penalizes hesitation missed by SPL.
4) *Progress:* $[(d_0 - d_f)/d_0]_0^1$ for initial and final geodesic distances d_0 and d_f.
5) *HCR:* episodes with a human collision.
6) *NCR:* episodes colliding with a human seen for less than $1.0\,\mathrm{s}$ beforehand.
7) *PSC:* percentage of steps with at least $1.0\,\mathrm{m}$ clearance from every human, following Falcon [2].

C. Main Results in Simulation

Table I shows the cost of LOS-only reasoning. Even with ground-truth target coordinates, Falcon cannot anticipate humans behind walls. In hidden boundary, its NCR is 8.33% against our 5.65%; its HCR is 25.30% and SR 13.99%, versus 14.29% and 37.50% for SAVNav. In mixed home activity, Falcon reaches 38.99% HCR and 10.42% SR, whereas SAVNav reaches 13.39% and 39.29%. Because collision ends an episode, these intrusions directly reduce completion. This is not only a goal-finding failure: in fully LOS crowded social, Falcon still collides in 30.95% of episodes. SAVNav instead yields before emergence using the active-node risk belief in (5).

ENMuS^3 exposes a different failure. Trained to isolate one target from interfering noise, it cannot separate the target from concurrent human sounds, yielding 0% SR and Progress $\leq 18.1\%$. Its high PSC is a by-product of wandering in open space and rarely approaching people. SAVNav instead converts the same sounds into the speed field in (11). It reaches targets while keeping PSC above 94% in every scenario, treating sound as a boundary rather than only a goal.

Reactive SAVNav shows the value of persistent map-level state. Removing $P_{\mathrm{tgt}}^{(t)}$ and $\mathbf{B}_{\mathrm{anch}}^{(t)}$ delays target resolution and costs about 15 SR points per scenario; SPL and STL fall accordingly. Its hidden-boundary HCR is the lowest at 13.39%, but this reflects timid progress rather than effective completion.

D. Ablation Study

Without topology-aware acoustic anticipation, only LOS sources raise social cost; an NLOS human matters only after entering view. In hidden boundary, HCR rises from 14.29% to 19.05%, NCR from 5.65% to 6.85%, and SR falls to 34.23%. Blind-corner safety therefore requires predicting where risk emerges rather than waiting for appearance. The mixed scenario narrows the gap because its LOS group dominates, although the full system remains ahead on every metric.

Without active sensory exploration, the robot drops the target belief map and disables (10). It greedily selects the nearest traversable point along the DOA while the target is unanchored. SR falls to 36.01% in both NLOS scenarios, with lower SPL and STL but little HCR change. The module thus improves target resolution and efficiency rather than safety.

E. Qualitative Analysis in Simulation

Fig. 5 links outcomes to internal state. In crowded social, SAVNav anchors `television` to its static object while `conversation` and `footsteps` bend the path around both the group and crossing pedestrian. In hidden boundary, `footsteps` raises the junction risk belief (5). Its anisotropic cost field (7) covers the junction before the pedestrian appears at step 185, so the path detours before visual evidence. In mixed home activity, active exploration selects a viewpoint via (10) at step 103. After anchoring, the conversational-group cost keeps the path clear at step 265. The ablations fail complementarily: one misses the NLOS caller, while the other collides at the junction.

F. Real-World Experiments

1) Setup: We deploy SAVNav in a $6.50 \times 7.45\,\mathrm{m}$ room with $33.95\,\mathrm{m}^2$ traversable area and hidden-boundary topology. A head-mounted RealSense D435if and ReSpeaker 4-channel array provide observations; an external RealSense D435i supplies ground-truth DOA per active source. Loudspeakers replay dataset recordings, and actors form the conversational group and dynamic pedestrian. Onboard audio performs sound-event classification only. We run six episodes per scenario with the simulation metrics.

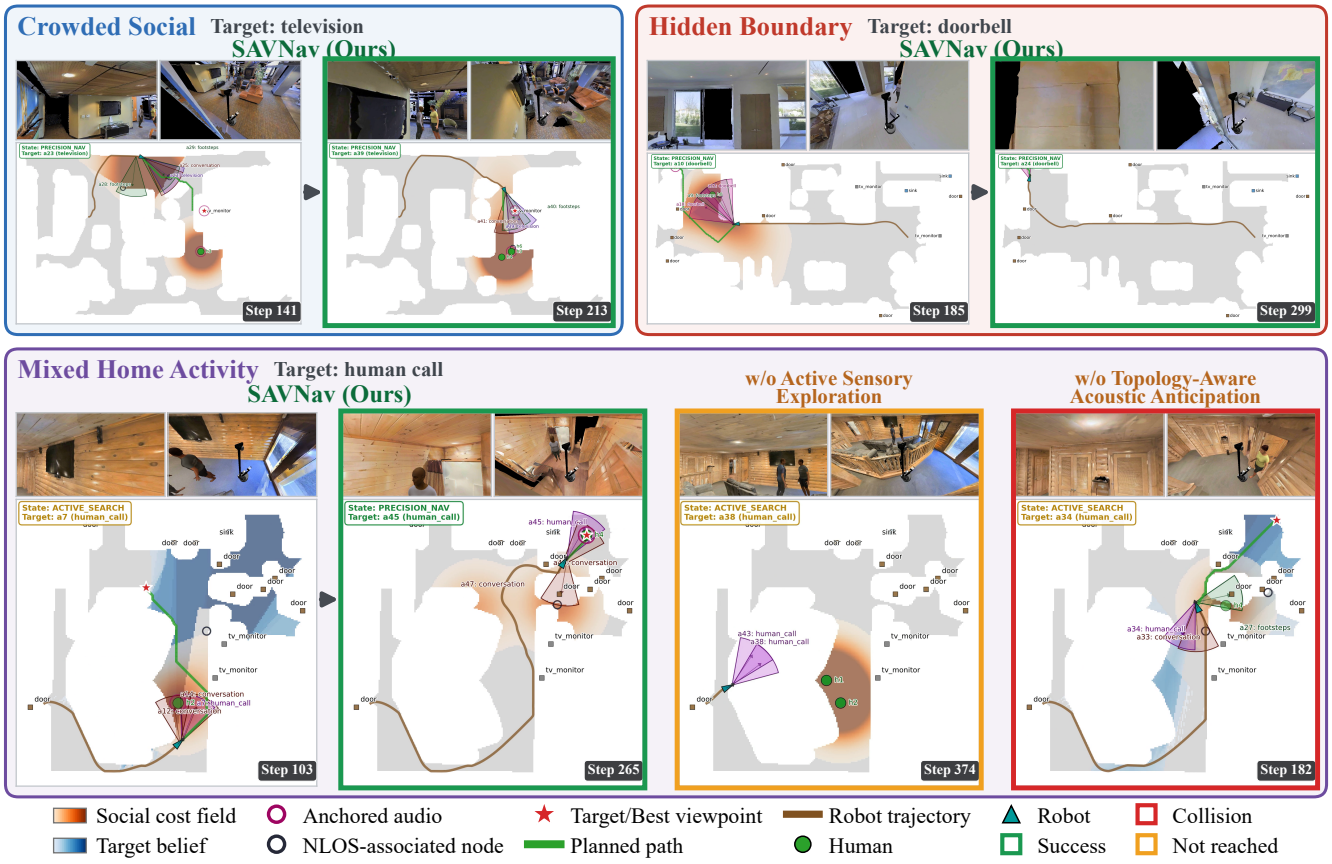


Fig. 5. Simulation examples for crowded social, hidden boundary, and mixed home activity. Green, red, and orange borders mark success, collision, and failure to reach the target. Each frame pairs camera views with maps showing the social cost field, target belief, anchored audio, non-line-of-sight (NLOS)-associated nodes, planned path, and executed trajectory. The mixed-scene variants remove active sensory exploration or topology-aware acoustic anticipation as labeled.

TABLE II

PHYSICAL-ROBOT RESULTS: 18 EPISODES, SIX PER SCENARIO (PERCENT; ↑/↓: HIGHER/LOWER IS BETTER), USING GROUND-TRUTH DIRECTION OF ARRIVAL (DOA) PER SOURCE. METRICS ARE SUCCESS (SR), PATH/TIME EFFICIENCY (SPL/STL), PROGRESS, HUMAN/NLOS COLLISION RATES (HCR/NCR), AND PERSONAL-SPACE COMPLIANCE (PSC).

Scenario	SR↑	SPL↑	STL↑	Prog.↑	HCR↓	NCR↓	PSC↑
Crowded Social	66.67	49.57	20.85	81.68	0.00	0.00	93.15
Hidden Boundary	50.00	28.90	13.70	65.80	50.00	50.00	85.97
Mixed Home Activity	33.33	21.97	8.98	59.33	50.00	50.00	82.30

2) Results: Table II closes the loop on hardware without retuning. In crowded social, the robot attains 66.67% SR, no collision, and 93.15% PSC, showing that the anchored cost field survives real perception. The NLOS scenarios retain 65.80% and 59.33% Progress, so the target is still localized and approached. However, half their episodes collide with the emerging pedestrian, capping SR at 50.00% and 33.33%. Every hardware collision is therefore an NLOS one. Physical latency and footprint leave less avoidance margin than simulation. These staged trials are not directly comparable to the simulation set; they establish that anticipation, cost construction, and planning run together in real time.

3) Qualitative Analysis: In Fig. 6, the seated group and crossing pedestrian raise a cost field over their occupied region. The planned paths bend around it toward the doorbell in the first episode and the television in the second. The executed trajectories preserve this clearance to termination, showing the same mechanism as Fig. 5 under real audio and depth streams.

VII. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

SAVNav treats social sounds as spatial social boundaries while navigating to an acoustic target. SAVMap converts transient audio into persistent spatial constraints, and the training-free SAVNav policy couples active sensory exploration with socially-compliant navigation. In simulation, the framework reduces NLOS collisions and improves PSC over vision-based baselines. It then runs without retuning on a physical robot. Current limitations are the pre-computed semantic and topological map, limited acoustic interference in simulation, and too few physical trials for method comparison. Future work will jointly build topological and acoustic maps, evaluate stronger acoustic noise, and expand the real-world study.

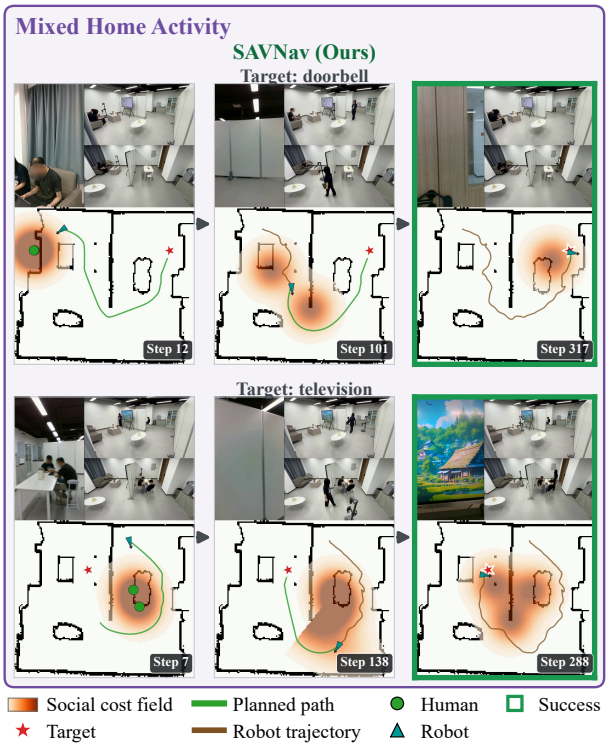


Fig. 6. Two successful physical-robot episodes, read left to right. Each frame pairs camera views with the social cost field, detected humans, planned path, and executed trajectory; green borders mark termination.

REFERENCES

[1] A. Francis, C. Pérez-D'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra, H.-T. L. Chiang, M. Everett, S. Ha, J. Hart, J. P. How, H. Karnan, T.-W. E. Lee, L. J. Manso, R. Mirsky, S. Pirk, P. T. Singamaneni, P. Stone, A. V. Taylor, P. Trautman, N. Tsoi, M. Vázquez, X. Xiao, P. Xu, N. Yokoyama, A. Toshev, and R. Martín-Martín, "Principles and guidelines for evaluating social robot navigation algorithms," *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 2, pp. 1–65, Jun. 2025.

[2] Z. Gong, T. Hu, R. Qiu, and J. Liang, "From cognition to precognition: A future-aware framework for social navigation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. Atlanta, GA, USA: IEEE, May 2025, pp. 9122–9129.

[3] Z. Fang, A. Xiao, D. Hsu, and G. H. Lee, "From Obstacles to Etiquette: Robot Social Navigation With VLM-Informed Path Selection," *IEEE Robotics and Automation Letters*, vol. 11, no. 4, pp. 3947–3954, Apr. 2026.

[4] Y.-J. Mun, M. Itkina, S. Liu, and K. Driggs-Campbell, "Occlusion-aware crowd navigation using people as sensors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. London, United Kingdom: IEEE, May 2023, pp. 12 031–12 037.

[5] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. Paris, France: IEEE, May 2020, pp. 9701–9707.

[6] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 15 511–15 520.

[7] S. Paul, A. Roy-Chowdhury, and A. Cherian, "AVLEN: Audio-visual-language embodied navigation in 3D environments," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 6236–6249.

[8] A. Younes, D. Honerkamp, T. Welschehold, and A. Valada, "Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 928–935, Feb. 2023.

[9] Z. Yang, J. Lin, P. Chen, A. Cherian, T. K. Marks, J. L. Roux, and C. Gan, "RILA: Reflective and imaginative language agent for zero-shot semantic audio-visual navigation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, Jun. 2024, pp. 16 251–16 261.

[10] Z. Shi, L. Zhang, L. Li, and Y. Shen, "Towards audio-visual navigation in noisy environments: A large-scale benchmark dataset and an architecture considering multiple sound-sources," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 14, pp. 14 673–14 680, Apr. 2025.

[11] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5.

[12] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind one embedding space to bind them all," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 15 180–15 190.

[13] K. Shimada, K. Uchida, Y. Koyama, T. Shibuya, S. Takahashi, Y. Mitsufuji, and T. Kawahara, "Open-vocabulary sound event localization and detection with joint learning of CLAP embedding and activity-coupled cartesian DOA vector," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 2946–2960, 2025.

[14] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone, "3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans," in *Robotics: Science and Systems XVI*. Robotics: Science and Systems Foundation, Jul. 2020.

[15] E. Bartoli, D. Rotondi, B. He, P. Jensfelt, K. O. Arras, and I. Leite, "Social 3D scene graphs: Modeling human actions and relations for interactive service robots," 2025.

[16] J. Even, J. Furrer, Y. Morales, C. T. Ishi, and N. Hagita, "Probabilistic 3-D mapping of sound-emitting structures based on acoustic ray casting," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 333–345, Apr. 2017.

[17] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Multimodal spatial language maps for robot navigation and manipulation," *International Journal of Robotics Research*, p. 2783649251351658, Jul. 2025.

[18] J. F. Mullen, D. Kumar, X. Qi, R. Madhivanan, A. Sen, D. Manocha, and R. Kim, "HomeEmergency - using audio to find and respond to emergencies in the home," *IEEE Robotics and Automation Letters*, vol. 10, no. 6, pp. 5649–5656, Jun. 2025.

[19] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee, "YOLO26: Key architectural enhancements and performance benchmarking for real-time object detection," Jan. 2026.

[20] J. A. Sethian, "Fast marching methods," *SIAM Review*, vol. 41, no. 2, pp. 199–235, Jan. 1999.

[21] X. Puig, E. Undersander, A. Szot, M. D. Cote, T.-Y. Yang, R. Partsey, R. Desai, A. Clegg, M. Hlavac, S. Y. Min, V. Vondruš, T. Gervet, V.-P. Berges, J. M. Turner, O. Maksymets, Z. Kira, M. Kalakrishnan, J. Malik, D. S. Chaplot, U. Jain, D. Batra, A. Rai, and R. Mottaghi, "Habitat 3.0: A Co-habitat for humans, avatars, and robots," in *The Twelfth International Conference on Learning Representations*, Oct. 2023.

[22] C. Chen, C. Schissler, S. Garg, P. Kobernik, A. Clegg, P. Calamia, D. Batra, P. Robinson, and K. Grauman, "SoundSpaces 2.0: A simulation platform for visual-acoustic learning," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 8896–8911.

[23] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, Mar. 2017, pp. 776–780.

[24] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[25] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D Data in Indoor Environments," in *2017 International Conference on 3D Vision (3DV)*, Oct. 2017, pp. 667–676.