ACOUSTIC WORLD MODELS FOR ACTIVE SCENE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Room impulse responses provide an indirect acoustic probe of scene geometry: as an agent moves, the recorded reverberation changes with nearby walls, openings, and free space. However, converting acoustic observations into maps is inherently ambiguous, since different reflector configurations can produce similar responses, especially when observations are sparse or motion is limited. We study active acoustic scene reconstruction, where an agent chooses sensing poses to improve its geometric belief. We introduce a cross-modal acoustic world model that encodes histories of RIRs and known poses into a motion-conditioned latent state used for both local occupancy decoding and future acoustic-latent prediction. At test time, candidate trajectories are rolled out in latent space, decoded into imagined occupancy maps, and scored using predicted map-space information gain and collision checks. We construct a synthetic benchmark of paired acoustic trajectories, poses, and floor-plan geometry. Experiments show improved local reconstruction over geometric and passive learned baselines. In closed-loop mapping, predictive rollout maintains on-par reconstruction while reducing the mean collision rate by approximately 80% relative to a matched CEM control. Its safety benefit also transfers to frontier exploration, with reduced coverage. Evaluation on held-out layouts and under additive measurement noise examines geometric transfer and acoustic robustness. Code and datasets will be made available upon publication.

1 INTRODUCTION

Autonomous agents require spatial beliefs to navigate, explore, and act safely, yet mature mapping systems often rely on sensors that can fail in challenging environments. Visual and visual-inertial systems Campos et al. (2021), Qin et al. (2018) and LiDAR or LiDAR-inertial systems Zhang et al. (2014), Shan et al. (2020), Xu et al. (2022) have enabled highly effective mapping pipelines, but cameras degrade under low illumination, smoke, occlusion, and transparent or reflective surfaces Bijelic et al. (2020), Quan et al. (2019). LiDAR and depth sensors can suffer from sparsity, specular failures, or limited coverage Yurtsever et al. (2020), Lötscher et al. (2023), Sajjan et al. (2020). GPS is often unreliable indoors Chang et al. (2023). Conventional perception channels may be vulnerable to physical adversarial attacks Nemcovsky et al. (2022), Zhang et al. (2021), Chen et al. (2024). These limitations motivate complementary sensing modalities for navigation and mapping Raza et al. (2023), Barnes et al. (2019), Younes et al. (2023).

Acoustic sensing is appealing because sound is independent of illumination and interacts with the scene through propagation, reflection, diffraction, and reverberation. In acoustic mapping, this interaction is often measured through room impulse responses (RIRs): temporal responses recorded at a listener after an acoustic impulse is emitted from a source. Under a fixed acoustic regime, RIRs depend on source-listener pose, room geometry, and material properties, providing indirect access to scene structure. However, acoustic geometry inference is severely underdetermined: a single RIR mixes direct sound, early reflections, late reverberation, and attenuation, constraining possible reflectors without uniquely identifying the surfaces that produced them Dokmanić et al. (2013), Antonacci et al. (2012), Moore et al. (2013). This ambiguity is strongest for a minimal omnidirectional source-receiver pair, where symmetries and overlapping reflection paths can make distinct geometries acoustically indistinguishable from one pose. We formulate acoustic scene reconstruction as an active, action-conditioned mapping problem. Existing acoustic methods often reduce ambiguity using specialized hardware, multiple microphones or sources, predefined scan

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.

trajectories, noise shaping, or geometric priors Dokmanić et al. (2013), Antonacci et al. (2012), Christensen et al. (2020), Yeon et al. (2024), Yang & Choudhury (2024), Serussi et al. (2024). In RIR-based mapping, each new measurement entangles pose, multipath propagation, and scene geometry. Its value therefore depends on the current belief and the measurements collected so far. The agent should choose poses expected to reduce this uncertainty.

Latent world models provide a natural framework for such predictive belief states. By learning action-conditioned latent dynamics, they allow agents to imagine future observations and evaluate candidate trajectories before acting Ha & Schmidhuber (2018), Hafner et al. (2019, 2023). Predictive representation methods, including JEPA-style models LeCun et al. (2022), Assran et al. (2025), further suggest that predicting abstract latent states rather than raw observations can emphasize temporally stable structure over high-frequency details. However, existing world models have primarily been developed for visual, proprioceptive, or simulated control domains. Acoustic mapping is different, with one-dimensional, indirect, multipath-dominated observations and spatial geometry as the desired output. We introduce a cross-modal acoustic world model for active scene reconstruction. Given a history of RIRs and known poses, the model learns a motion-conditioned latent state that decodes into local occupancy maps and predicts future acoustic latents under candidate motions. At test time, candidate trajectories are rolled out in latent space, decoded into imagined occupancy maps, and scored by predicted map-space entropy reduction and collision checks. The imagined maps expose how the occupancy belief may evolve along a candidate path before execution, supporting both information seeking and anticipatory obstacle avoidance. Since training requires trajectory-level supervision not provided by existing acoustic mapping datasets, we construct a synthetic benchmark from complementary floor-plan sources.

This paper makes four contributions. First, we introduce a geometry-grounded acoustic world model that combines motion-conditioned local reconstruction with future acoustic-latent prediction. Second, we develop an active mapping framework that evaluates candidate sensing trajectories through imagined occupancy beliefs. Third, we provide planner controls that separate predictive lookahead from the CEM optimizer: reconstruction remains on par while mean collision rate decreases, and the safety benefit transfers to frontier exploration. Fourth, we construct a large-scale benchmark for active acoustic mapping, comprising spatially indexed RIRs and metric geometry across complementary layout domains, with analyses of geometric diversity, acoustic variation, and robustness to measurement noise.

2 Related Work

Acoustic geometry inference and mapping. Classical methods in acoustic scene understanding recover room structure from RIRs using early reflections, time-of-arrival estimates, image-source models, and geometric fitting Dokmanić et al. (2013), Antonacci et al. (2012), Moore et al. (2013). Recent learning-based approaches map acoustic measurements to spatial representations: BatVision uses binaural cues for depth estimation Christensen et al. (2020), EchoScan predicts occupancy from array-based echoes Yeon et al. (2024), and MapLearn reconstructs structured indoor maps from acoustic trajectories Yang & Choudhury (2024). Related audio-visual navigation work demonstrates that sound supports embodied decision making, but typically targets navigation toward sound sources rather than reconstructing the geometry that produced the acoustic response Chen et al. (2020, 2022). These methods motivate belief-guided active mapping from a minimal acoustic setup.

World models and predictive representations. Latent world models learn compact action-conditioned dynamics that allow agents to imagine future states before acting Ha & Schmidhuber (2018), Hafner et al. (2019, 2023). Predictive-representation methods such as JEPA and V-JEPA similarly avoid raw reconstruction by predicting abstract latent targets, encouraging temporally stable representations LeCun et al. (2022), Assran et al. (2025). Acoustic mapping differs from the visual and proprioceptive domains where these models are typically applied: reverberant audio is dominated by multipath propagation, and raw waveform prediction can overemphasize phase-sensitive details that are weakly aligned with geometry. Our cross-modal model predicts future acoustic latents while grounding the world state with an occupancy decoder, so rollouts capture how motion changes both sound and map belief.

Active mapping and information-driven exploration. Classical frontier exploration methods optimize motion toward boundaries between known free space and unexplored regions Yamauchi

(1997), often using graph search methods such as A* Hart et al. (1968). Learning-based approaches, including Active Neural SLAM and curiosity-driven exploration, optimize coverage or intrinsic novelty objectives Chaplot et al. (2020), Pathak et al. (2017). These methods work well when the current map is produced by relatively direct sensors. Acoustic observations, however, are indirect and ambiguous: the most informative action may not correspond to the nearest frontier, and the current occupancy belief may be diffuse or misleading. Our planner scores predicted occupancy evolution using an acoustic world model.

3 METHOD

Our method uses acoustic histories to estimate local geometry and predict the effect of future motion. Given recent monaural RIRs and known poses, the model forms a latent scene belief and decodes local occupancy. For a candidate motion, it predicts a future acoustic latent and updates the imagined belief. Repeating this process produces future maps for trajectory scoring. The agent executes the first action, receives a measured RIR, and replans. Monaural omnidirectional sensing provides a minimal setup in which spatial inference relies on acoustic evidence collected through motion.

3.1 ACTIVE ACOUSTIC MAPPING

Let $\mathcal{S}$ denote the physical acoustic scene and $\mathbb{A}$ the continuous motion action space of an agent. The reconstruction target is the scene's global binary occupancy projection

$$M^\star = \Pi_{\mathrm{occ}}(\mathcal{S}), \qquad M^\star \in \{0,1\}^{H_g \times W_g}, \tag{1}$$

where Π_{occ} rasterizes the scene into a binary wall-occupancy map. The ground-truth map $M^\star$ represents the wall/free-space structure induced by the floor-plan polygon. At time t, the agent is at pose $p_t \in \mathbb{R}^2$ and executes a continuous motion action $a_t \in \mathbb{A}$. The pose evolves according to a deterministic planar transition model that determines p_{t+1} given (p_t, a_t), with feasibility and collision handling described in Appendix G. We assume that p_t is available from odometry, simulation, or an external localization system. At each pose, the agent emits an acoustic impulse probe and records a single-channel RIR,

$$x_t \sim p(x \mid \mathcal{S}, p_t), \qquad x_t \in \mathbb{R}^{L_{\mathrm{rir}}}, \tag{2}$$

where L_{rir} is the number of waveform samples. The passive acoustic inverse problem is to recover the occupancy projection of the hidden scene from the observed RIRs and known poses:

$$\hat{M}_t = F_\theta(x_{1:t}, p_{1:t}) \approx \mathbb{E}\left[\Pi_{\mathrm{occ}}(\mathcal{S}) \mid x_{1:t}, p_{1:t}\right]. \tag{3}$$

This inverse is generally non-unique: different scenes or reflector configurations can induce similar RIRs at a given pose. We therefore condition reconstruction on the interaction history $h_t = (x_{1:t}, p_{1:t}, a_{1:t-1})$. $F_\theta(h_t)$ produces a global occupancy belief $\hat{M}_t \in [0,1]^{H_g \times W_g}$, where $\hat{M}_t(i,j)$ denotes the probability that global cell (i,j) is occupied. In the active setting, the agent must also choose actions that improve this belief under a finite sensing budget T_{ep}:

$$\pi^\star = \arg\min_\pi \mathbb{E}_\pi \left[\mathcal{L}_{\mathrm{map}}\left(\hat{M}_{T_{\mathrm{ep}}}, M^\star\right)\right], \tag{4}$$

where $\mathcal{L}_{\mathrm{map}}$ measures geometric reconstruction quality.

3.2 DATA

No public corpus provides paired RIR trajectories, source-receiver configurations, and the 2D floor-plan geometry required by our setting. We therefore use the geometric acoustic ray tracer of SoundSpaces 2.0 Chen et al. (2022) to construct a unified acoustic mapping corpus from three complementary layout sources. Each generated scene defines a physical acoustic environment $\mathcal{S}_s$: a floor polygon in a canonical metric frame, its extruded 3D geometry, a dense set of source-receiver configurations, and the corresponding simulated RIRs. The corpus contains 3,251 generated scenes and approximately 2.4M spatially indexed RIRs. Dense spatial sampling supports the construction of sensing trajectories and future-observation targets from each scene.

For each scene, we store configurations $\{(p_i^{\mathrm{src}}, p_i^{\mathrm{mic}}, x_i)\}_{i=1}^{N_s}$, where p_i^{src} and p_i^{mic} are source and receiver positions and x_i is the corresponding RIR. The model represents each configuration by its 2D

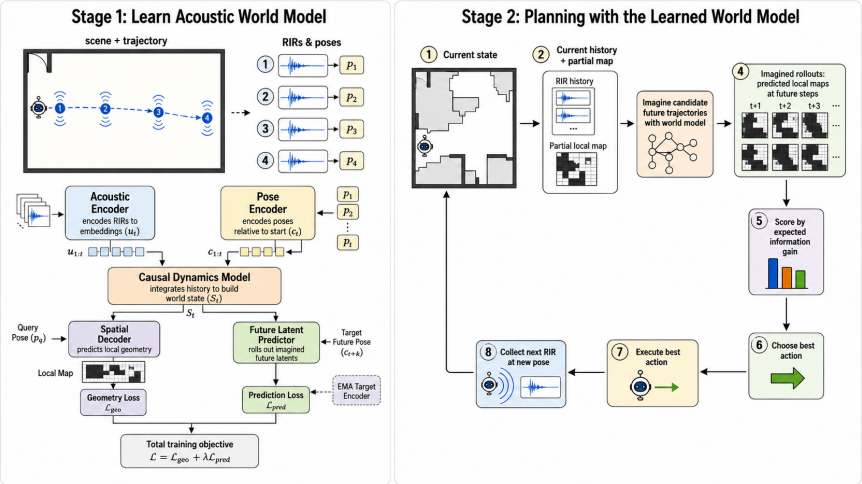


Figure 1: Overview of the proposed acoustic world model and active mapping pipeline. From a short history of RIRs and known poses, the model learns a motion-conditioned world state for local geometry decoding and future latent prediction. At test time, the planner rolls out candidate trajectories with the world model, decodes imagined occupancy maps, scores them by map-space information gain, and executes the first action in a receding-horizon loop.

sensing pose p_i. Local ego-centric occupancy targets $m_q^{\ell\star}$ are rasterized around the query pose from the metric floor polygon (Equation (7)). The same corpus supports fixed-trajectory reconstruction, motion-conditioned future prediction, and closed-loop active mapping. Scene-level splits keep training and test wall geometries disjoint.

Matterport3D Chang et al. (2017) provides single-room layouts derived from panoramic captures using the annotations of Zou et al. (2021). **AtlantaLayout** Pintore et al. (2020) contributes non-Manhattan single-room layouts. **ResPlan** Abouagour & Garyfallidis (2025) provides larger residential floor plans and is held out for active mapping evaluation. We use its outer apartment boundary as one acoustic envelope, omitting internal subdivisions. The sources span different scales, shapes, and boundary complexity under a shared simulator and material configuration. Geometry induces shifts in early decay, clarity, center time, and direct-to-reverberant ratio despite similar late reverberation times. Appendix A reports construction details, generated-corpus and experimental split counts, and geometric and acoustic analyses.

3.3 Acoustic world model

The acoustic world model maps recent RIRs and poses into latent belief tokens for local occupancy decoding and future acoustic prediction. The mapping backend accumulates local estimates in the global frame (Section 3.4). Architecture and training details are in Appendix C.

Relative coordinate convention. The model uses a sliding context of length $K \leq K_{\max}$. For a window ending at t, poses are expressed relative to the first pose $\tilde{p}_i = p_i - p_{t-K+1}$, where $i \in \{t-K+1, \ldots, t\}$. This makes the representation invariant to a common translation of the sensing trajectory. We suppress direct-path RIR samples before encoding and represent the omnidirectional source-receiver pair by its 2D position. Relative poses are embedded with random Fourier features Tancik et al. (2020) followed by an MLP:

$$c_i = \rho_\theta\left([\sin(2\pi\tilde{p}_i B), \cos(2\pi\tilde{p}_i B)]\right) \in \mathbb{R}^D, \tag{5}$$

where $B \in \mathbb{R}^{2\times m}$ is a fixed Gaussian random-feature matrix, m is the number of sampled frequencies, and ρ_θ is the MLP producing the pose embedding c_i.

Blind acoustic encoder. Each RIR is first encoded independently by a "blind" acoustic encoder $u_i = e_\theta(x_i)$, where $u_i \in \mathbb{R}^D$. This encoder performs single-RIR-level feature extraction without receiving pose encoding, and so it is "blind" to trajectory context and room geometry. This separates pose-

agnostic acoustic features from motion-conditioned reasoning. It is implemented as a Transformer Encoder over complex-STFT frame tokens with temporal positional encodings.

Causal dynamics model. The acoustic embeddings $u_{t-K+1:t}$ from the blind encoder and the pose embeddings $c_{t-K+1:t}$ are integrated by a causal Transformer into trajectory-level latent state tokens,

$$S_t = \phi_\theta(u_{t-K+1:t}, c_{t-K+1:t}), \qquad S_t \in \mathbb{R}^{K\times D}. \tag{6}$$

The causal Transformer ϕ_θ preserves evidence from multiple acoustic viewpoints in S_t. Motion conditioning uses adaptive layer normalization (AdaLN) Peebles & Xie (2023), Maes et al. (2026).

Local occupancy decoder. The latent belief tokens are grounded in geometry by a local occupancy decoder. While Eq. (3) defines the inverse problem at the level of the global occupancy map $M^\star$, the neural model implements a local version of this inverse problem around a query pose p_q. We define the corresponding local target as

$$m_q^{\ell\star} = \Pi_{p_q}^{\ell}(M^\star), \qquad m_q^{\ell\star} \in \{0,1\}^{H_\ell\times W_\ell}, \tag{7}$$

where $\Pi_{p_q}^{\ell}$ extracts an ego-centric local occupancy crop centered at p_q. The local acoustic inverse estimator is

$$\hat{m}_q^{\ell} = d_\theta(S_t, \tilde{p}_q) \approx \mathbb{E}\left[m_q^{\ell\star} \mid h_t, p_q\right], \qquad \hat{m}_q^{\ell} \in [0,1]^{H_\ell\times W_\ell}. \tag{8}$$

During inference, the query pose is the last pose in the trajectory window, so the output is an ego-centric local map around the current agent position.

We decode local geometry because nearby acoustic evidence is more reliable Luo et al. (2022), while global consistency is accumulated by the mapping backend. The decoder initializes learnable tokens on a 2D local-map grid, adds cell-coordinate embeddings, cross-attends them to S_t, and maps the resulting per-cell features to occupancy logits.

To support planning, the model predicts the blind acoustic latent expected at a future pose: $\hat{u}_{t+k} = g_\theta(S_t, c_{t+k})$, $\hat{u}_{t+k} \in \mathbb{R}^D$. We predict latents rather than raw RIR waveforms because waveform prediction is high-dimensional and phase-sensitive, while the planning objective depends on geometry-relevant acoustic structure. This latent predictor is a Transformer decoder, with the implementation details brought in Section C.

Autoregressive latent rollout. For planning, future latents are generated autoregressively. Given S_t, the model predicts $\hat{u}_{t+1} = g_\theta(S_t, c_{t+1})$. The predicted latent is appended to the trajectory context, and the latent belief tokens (i.e. the world state) are updated:

$$\hat{S}_{t+1} = \phi_\theta\left([u_{t-K+2:t}, \hat{u}_{t+1}], [c_{t-K+2:t}, c_{t+1}]\right). \tag{9}$$

Once the context exceeds $K_{\max}$, the oldest element is dropped before recomputing the latent tokens. Repeating this process over a candidate future trajectory yields imagined latent token sequences $\hat{S}_{t+1:t+H_{\text{plan}}}$ and corresponding local occupancy predictions for candidate future trajectories $\hat{m}^{\ell}_{t+1:t+H_{\text{plan}}}$. Together, these modules instantiate the learned estimator F_θ from Eq. (3). We denote the full world-model implementation by $F_\theta = (e_\theta, \rho_\theta, \phi_\theta, d_\theta, g_\theta)$

Training objective. The model is trained on histories h_t, query poses p_q, local occupancy targets $m_q^{\ell\star}$, and future RIRs x_{t+k}. The objective is

$$\mathcal{L} = \mathcal{L}_{\text{geo}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}},$$

with local geometry supervision

$$\mathcal{L}_{\text{geo}} = \lambda_{\text{focal}}\mathcal{L}_{\text{focal}}(\hat{m}_q^{\ell}, m_q^{\ell\star}) + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}(\hat{m}_q^{\ell}, m_q^{\ell\star}),$$

and future latent prediction

$$\mathcal{L}_{\text{pred}} = \left\|\hat{u}_{t+k} - \text{sg}\left(e_{\bar{\theta}}(x_{t+k})\right)\right\|_2^2.$$

Here sg stops gradients through the target, and $e_{\bar{\theta}}$ is an exponential-moving-average (EMA) copy of the blind acoustic encoder. Both $\hat{u}_{t+k}$ and $e_{\bar{\theta}}(x_{t+k})$ lie in the same pose-agnostic acoustic latent space. During planning, unavailable future RIRs are replaced by autoregressively predicted latents.

3.4 PREDICTIVE PLANNING

The neural model predicts local occupancy, but active mapping requires a global belief accumulated over the full trajectory. We maintain a global log-odds map $L_t \in \mathbb{R}^{H_g \times W_g}$, with occupancy probabilities $P_t = \sigma(L_t)$. After receiving an observed RIR at pose p_t, the model predicts a local map $\hat{m}_t^\ell$, which is inserted into the global frame by a placement operator $\mathcal{W}_{p_t}$:

$$L_t = L_{t-1} + \mathcal{W}_{p_t}\left(\operatorname{logit}\left(\operatorname{clip}\left(\hat{m}_t^\ell, \epsilon, 1-\epsilon\right)\right)\right), \tag{10}$$

where ϵ is a small numerical constant. The learned model therefore provides local acoustic evidence, while log-odds fusion accumulates it into a global occupancy belief. At test time, the learned world model is used for receding-horizon planning. At each decision step, the planner treats the world model as an internal acoustic simulator. It samples candidate future action sequences:

$$A_{t:t+H_{\text{plan}}-1}^{(n)} = \left(a_t^{(n)}, \ldots, a_{t+H_{\text{plan}}-1}^{(n)}\right), \qquad n = 1, \ldots, N. \tag{11}$$

Each sequence induces a candidate future pose trajectory. For each candidate, the model performs a deterministic imagined latent rollout: future acoustic latents are autoregressively predicted, local occupancy maps are decoded, and the imagined local maps are fused into a temporary future belief:

$$\tilde{L}_{t+H_{\text{plan}}}^{(n)} = \text{ImagineFuse}\left(L_t, h_t, A_{t:t+H_{\text{plan}}-1}^{(n)}, F_\theta\right), \qquad \tilde{P}_{t+H_{\text{plan}}}^{(n)} = \sigma\left(\tilde{L}_{t+H_{\text{plan}}}^{(n)}\right). \tag{12}$$

The operator ImagineFuse denotes the rollout-and-fusion procedure induced by the current history, candidate action sequence, learned world model, and occupancy fusion rule. Implementation details are given in Appendix G.

We score each candidate using a plug-in map-space information-gain surrogate: the entropy reduction induced by the model's deterministic imagined rollout. Let $\mathcal{H}(p) = -p\log p - (1-p)\log(1-p)$, denote Bernoulli entropy. The score of candidate $A^{(n)}$ is

$$R(A^{(n)}) = \sum_{i,j}\left[\mathcal{H}(P_t(i,j)) - \mathcal{H}\left(\tilde{P}_{t+H_{\text{plan}}}^{(n)}(i,j)\right)\right]. \tag{13}$$

This deterministic score estimates the reduction in occupancy uncertainty under the predicted future observations.

We optimize this score with CEM Rubinstein (1997): candidate action sequences are sampled, scored by Eq. (13), and the sampling distribution is refit to elite trajectories. The full score includes collision penalties in the imagined map and a boundary-discovery bonus. Predictive occupancy supplies the geometry used to assess future collisions. The agent executes only the first action, observes a real RIR, updates its history and map, and replans. Feasibility filtering, collision handling, and CEM hyperparameters are given in Appendix G.

4 RESULTS

We evaluate local reconstruction from fixed acoustic histories (Section 4.1) and closed-loop mapping on held-out layouts (Section 4.2). The first tests motion-conditioned acoustic inference. The second tests how predictive planning affects reconstruction, exploration, and collisions under a fixed sensing budget. Appendix L examines the learned latent representation.

4.1 LOCAL RECONSTRUCTION

Task. Each sample is a context window $h_t^{(K)} = (x_{t-K+1:t}, p_{t-K+1:t})$, where x_i is a single-channel RIR and p_i is the corresponding 2D sensing position. Unless otherwise stated, we use $K = 10$ observations. The effect of context length is analyzed in Appendix H. The model outputs a local occupancy probability map $\hat{M}_t^{\text{loc}} \in [0,1]^{H_\ell \times W_\ell}$, covering a 10×10m region centered at the final pose. Full target construction and augmentation details are provided in Appendix C.

We evaluate local reconstruction on the room-scale subsets of the synthesized corpus described in Section 3.2: Matterport3D Chang et al. (2017), Zou et al. (2021) and AtlantaLayout Pintore et al. (2020). These datasets provide training, validation, and matched-distribution test scenes whose scale

makes short acoustic histories informative for local occupancy estimation. Full dataset statistics and construction details are provided in Appendix A, and architecture and training hyperparameters in Appendix C.

Baselines and ablations. We compare our method to geometric and learned-acoustic baselines. The **Convex hull** and **Alpha shape** Edelsbrunner et al. (2003) baselines are geometry-only heuristics based on the microphone trajectory Preparata & Shamos (1985). They quantify the geometry recoverable from the sensing trajectory alone. The **EchoScan-style passive predictor** is a learned acoustic-to-occupancy baseline inspired by EchoScan Yeon et al. (2024). Since EchoScan does not release code and assumes a different sensing regime, we adapt its CNN mapping principle to our single-sensor moving-agent setting, using multiple relative-pose RIRs to predict a global floorplan without trajectory-level dynamics. Other acoustic mapping methods Dokmanić et al. (2013), Kreković et al. (2016), Christensen et al. (2020) require different sensing configurations or geometric assumptions. Appendix E details the baseline adaptations and evaluation protocol.

The geometric baselines use $K = 50$ poses, while both learned models use $K = 10$ RIRs and poses. The adapted EchoScan model was tuned on validation data and trained until validation performance showed strongly diminishing gains, with native-output learning curves in Appendix E.2.

We report Chamfer distance between thresholded predictions and skeletonized ground-truth walls, tolerance-based ϵ-F1 over boundary points, and distance-field mean absolute error (DF-MAE), computed from Euclidean distance transforms to the nearest wall. Full metric definitions are provided in Appendix D.

Table 1: Quantitative evaluation of local acoustic-to-geometry reconstruction. All threshold-dependent metrics are reported using the operating threshold selected independently on the validation set. Values are reported as mean $\pm$ standard deviation.

Method	**Geometric Error**		**ϵ-F1**		
	Chamfer (m) $\downarrow$	**DF-MAE (m)** $\downarrow$	$F1_{\epsilon=0.15\mathrm{m}} \uparrow$	$F1_{\epsilon=0.30\mathrm{m}} \uparrow$	$F1_{\epsilon=0.45\mathrm{m}} \uparrow$
Alpha Shape ($K = 50$)	0.485 ± 0.308	0.391 ± 0.276	0.412 ± 0.127	0.582 ± 0.164	0.714 ± 0.186
Convex Hull ($K = 50$)	0.483 ± 0.293	0.382 ± 0.249	0.416 ± 0.122	0.587 ± 0.155	0.719 ± 0.175
EchoScan-style ($K = 10$)	0.565 ± 0.436	0.371 ± 0.321	0.488 ± 0.190	0.657 ± 0.227	0.736 ± 0.210
Ours ($K = 10$)	$\mathbf{0.150 \pm 0.084}$	$\mathbf{0.110 \pm 0.095}$	$\mathbf{0.799 \pm 0.096}$	$\mathbf{0.947 \pm 0.083}$	$\mathbf{0.970 \pm 0.063}$

Table 1 shows improved local reconstruction across all reported metrics. The gains over trajectory-only heuristics support the contribution of acoustic evidence, while the comparison with the adapted passive predictor supports motion-conditioned modeling. Appendix F additionally reports an additive-noise robustness evaluation without model retraining. Performance is nearly unchanged at 20 and 30 dB SNR and degrades moderately at 10 dB, providing a controlled test of robustness to imperfect acoustic measurements.

4.2 Active acoustic mapping

We next evaluate the full closed-loop active mapping system on the held-out ResPlan Abouagour & Garyfallidis (2025) scenes, whose larger whole-apartment envelopes make sequential exploration and planning meaningful. Dataset statistics are provided in Appendix A. This setup tests whether the learned acoustic world model is useful for decision making beyond local reconstruction from fixed trajectories. Each episode starts from a valid pose p_1 in an unseen test scene. After an initial random warm-up context $h_K = (x_{1:K}, p_{1:K}, a_{1:K-1})$, the planner controls the agent for a fixed budget of T steps. At each step t, the agent executes an action a_t, records a new RIR x_{t+1}, predicts a local occupancy map $\hat{m}^{\ell}_{t+1}$, fuses it into the global log-odds belief L_{t+1}, and selects the next sensing action. All planners use the same learned local acoustic mapper d_θ and global Bayesian fusion backend. Planners differ only in how a_t is selected. Additional implementation details are provided in Appendices C.4 and G.

Planners. We compare Random Walk, a reactive DFS heuristic, and classical Frontier + A* Yamauchi (1997), Hart et al. (1968). Two planner controls assess predictive lookahead. CEM (No-WM) retains the optimizer, horizon, current-map rejection rules, and fallback, and scores frontier proximity in the current belief. Frontier + A* (WM) adds predicted occupancy checks to frontier navigation. CEM (WM) jointly evaluates predicted information gain and safety. A GT-occupancy oracle replaces

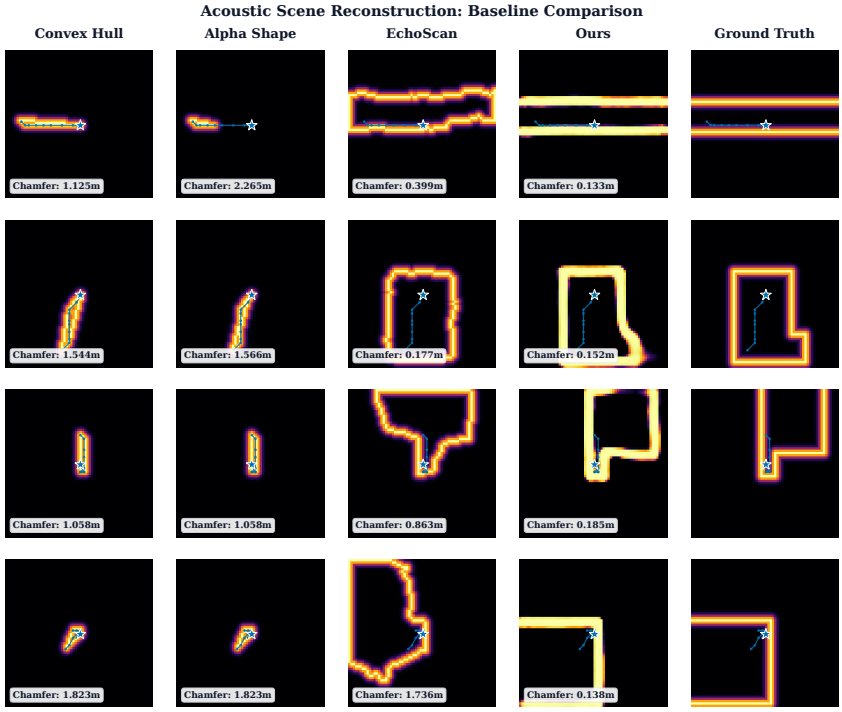


Figure 2: Local reconstruction on held-out scenes. Learned methods use $K = 10$ observations and geometric baselines use $K = 50$ poses. Columns compare predictions with ground truth.

Table 2: Closed-loop active acoustic mapping under a fixed budget of 100 environment steps. Values are mean $\pm$ standard deviation across test episodes. CD and DF denote Chamfer and DF-MAE (m), SER is in m^2/m, and crash rate is per 100 m. Bold and underlining mark the best and second-best non-oracle means.

Planner	**Exploration / Stability**			**Geometric Error**		**ϵ-F1**		
	Cov.% ↑	**SER ↑**	**Crash ↓**	**CD ↓**	**DF ↓**	**0.15 m ↑**	**0.30 m ↑**	**0.45 m ↑**
Random Walk	58.27 ± 17.12	3.50 ± 0.40	4.94 ± 8.47	0.345 ± 0.128	0.348 ± 0.196	0.254 ± 0.040	0.589 ± 0.075	0.766 ± 0.082
DFS Heuristic	75.16 ± 22.68	$\underline{11.96 \pm 14.98}$	$\mathbf{0.32 \pm 3.07}$	0.373 ± 0.149	0.339 ± 0.144	0.268 ± 0.043	0.617 ± 0.085	0.790 ± 0.092
Frontier + A*	92.71 ± 13.89	11.46 ± 11.89	22.55 ± 90.18	0.264 ± 0.108	0.204 ± 0.108	$\underline{0.295 \pm 0.039}$	$\underline{0.682 \pm 0.074}$	$\underline{0.863 \pm 0.076}$
Frontier + A* (WM)	84.32 ± 19.63	$\mathbf{15.95 \pm 18.76}$	1.18 ± 4.73	0.322 ± 0.126	0.251 ± 0.123	0.268 ± 0.045	0.624 ± 0.087	0.805 ± 0.091
CEM (No-WM)	$\mathbf{95.94 \pm 9.25}$	6.18 ± 7.83	3.89 ± 34.99	$\underline{0.254 \pm 0.072}$	$\mathbf{0.182 \pm 0.073}$	0.285 ± 0.028	0.666 ± 0.052	0.854 ± 0.054
CEM (WM)	$\underline{93.65 \pm 11.76}$	7.38 ± 7.73	$\underline{0.78 \pm 3.21}$	$\mathbf{0.253 \pm 0.083}$	$\underline{0.196 \pm 0.085}$	$\mathbf{0.299 \pm 0.033}$	$\mathbf{0.691 \pm 0.064}$	$\mathbf{0.869 \pm 0.065}$
GT-occupancy oracle	96.63 ± 9.78	7.45 ± 7.44	0.28 ± 1.69	0.238 ± 0.070	0.187 ± 0.082	0.304 ± 0.031	0.701 ± 0.059	0.880 ± 0.059

predicted future maps with ground-truth local occupancy during candidate scoring. It provides a reference for planning with accurate future geometry. Full definitions are in Appendix G.

Metrics. We report exploration, safety, and reconstruction metrics. Coverage measures the fraction of ground-truth wall cells that fall inside the explored region, while Spatial Efficiency Ratio (SER) measures explored area per meter traveled. Safety is measured by crash rate, reported as collisions per 100 meters. Reconstruction quality is evaluated on the final global map within the explored region using Chamfer distance, tolerance-based ϵ-F1, and DF-MAE, as in Section 4.1. Threshold-dependent metrics use an occupancy threshold of 0.5. Full metric definitions are provided in Appendix D.

Table 2 evaluates safety jointly with mapping performance. DFS attains the lowest non-oracle mean collision rate, with substantially lower coverage and worse geometric reconstruction. The matched CEM comparison separates predictive lookahead from the optimizer. Relative to CEM (No-WM), CEM (WM) maintains on-par reconstruction while reducing mean collision rate from 3.89 to 0.78 collisions per 100 m, a reduction of approximately 80%. Chamfer remains essentially unchanged (0.254 versus 0.253 m), all three F1 scores improve, and DF-MAE changes from 0.182 to 0.196 m. Coverage decreases from 95.94% to 93.65%, while SER increases from 6.18 to 7.38. These are descriptive comparisons of episode means, with the collision-rate variability reported in the table.

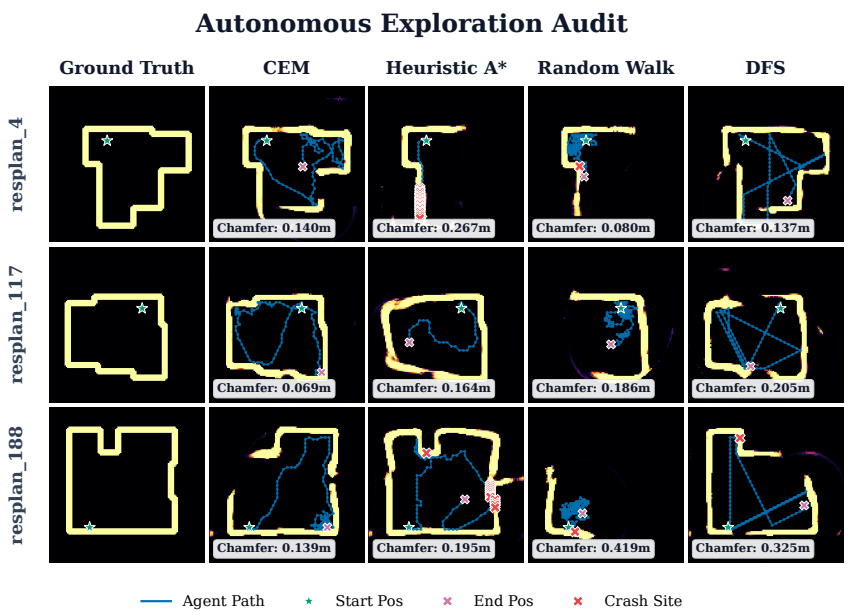


Figure 3: Autonomous exploration across planners. We show final reconstructed maps and trajectories comparing our proposed CEM (WM) against the myopic Frontier + A* (without WM) in identical unknown acoustic environments. Rows show the final fused Bayesian prediction overlaid with the executed trajectory and Chamfer distance at $\tau = 0.5$.

Predictive lookahead also transfers to Frontier + A*: mean collision rate decreases from 22.55 to 1.18 per 100 m, approximately 95%. Coverage decreases from 92.71% to 84.32%, with reduced reconstruction quality, indicating a more conservative exploration policy. Joint CEM optimization better preserves exploration and reconstruction while using predicted occupancy for safety. Figure 3 illustrates how errors in the current map can lead to repeated attempts to traverse blocked space.

Across the reported criteria, CEM (WM) generally trends toward the GT-occupancy oracle reference, with the clearest remaining gap in collision rate. This suggests scope for improving planning through more accurate future geometry prediction. Rollout-horizon analysis, additional budget experiments, and qualitative examples appear in Appendices I, M, and K.

5 Conclusion

We presented a cross-modal acoustic world model that uses motion-conditioned RIR histories for local geometry reconstruction and future acoustic-latent prediction. Local reconstruction improves over geometric and adapted passive learned baselines. In closed-loop mapping, predictive rollout maintains on-par reconstruction relative to a matched CEM control while substantially reducing mean collision rate. Its safety benefit also transfers to frontier exploration, with a more conservative exploration trade-off.

Evaluation covers diverse simulated layouts and additive measurement noise without model retraining. Training uses clean RIRs under a shared material configuration and empty architectural shells. Material variation, clutter, calibration errors, structured background sound, measured real-world RIRs, and noisy odometry remain untested. The agent assumes known poses, leaving localization and global pose correction to future work. Acoustic mapping may support navigation in visually degraded settings, while deployment requires attention to consent and indoor privacy.

RIR propagation is simulated in 3D, while reconstruction and motion are planar. A 3D extension requires volumetric or implicit occupancy decoding and fusion, with higher representation cost. Full 3D navigation also expands the action space and introduces vertical sensing constraints.

AI USE STATEMENT

Generative AI tools assisted with grammar, spelling and word-choice during writing the manuscript, and with the creation of some of the figures. We also used such tools to verify the correctness of our code.

REPRODUCIBILITY STATEMENT

Dataset construction and acoustic statistics are documented in Appendix A, architecture and training in Appendix C, metrics in Appendix D, and planner definitions in Appendix G. Appendices E.2 and F report baseline convergence and measurement-noise results. Source code, training and evaluation data, and the active mapping environment will be made publicly available upon acceptance.

REFERENCES

Mohamed Abouagour and Eleftherios Garyfallidis. ResPlan: A large-scale vector-graph dataset of 17,000 residential floor plans. *arXiv preprint arXiv:2508.14006*, 2025.

Fabio Antonacci, Jason Filos, Mark RP Thomas, Emanuël AP Habets, Augusto Sarti, Patrick A Naylor, and Stefano Tubaro. Inference of room geometry from acoustic impulse responses. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(10):2683–2695, 2012.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Dan Barnes, Rob Weston, and Ingmar Posner. Masking by moving: Learning distraction-free radar odometry from pose information. *arXiv preprint arXiv:1909.03752*, 2019.

Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11682–11692, 2020.

Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE transactions on robotics*, 37(6):1874–1890, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017.

Yingxiu Chang, Yongqiang Cheng, Umar Manzoor, and John Murray. A review of uav autonomous navigation in gps-denied environments. *Robotics and Autonomous Systems*, 170:104533, 2023.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155*, 2020.

Baodong Chen, Wei Wang, Pascal Sikorski, and Ting Zhu. Adversary is on the road: Attacks on visual {SLAM} using unnoticeable adversarial patch. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 6345–6362, 2024.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *European conference on computer vision*, pp. 17–36. Springer, 2020.

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 (NeurIPS) Datasets and Benchmarks Track*, 2022.

Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see 3d spatial layout with two ears. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1581–1587. IEEE, 2020.

Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M Lu, and Martin Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 110(30):12186–12191, 2013.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on information theory*, 29(4):551–559, 2003.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. doi: 10.1109/TSSC.1968.300136.

International Organization for Standardization. ISO 3382-1: Acoustics — measurement of room acoustic parameters — Part 1: Performance spaces, 2009. ISO Standard.

Miranda Kreković, Ivan Dokmanić, and Martin Vetterli. Echoslam: Simultaneous localization and mapping with acoustic echoes. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11–15. Ieee, 2016.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Michael Lötscher, Nicolas Baumann, Edoardo Ghignone, Andrea Ronco, and Michele Magno. Assessing the robustness of lidar, radar and depth cameras against ill-reflecting surfaces in autonomous vehicles: An experimental study. In *2023 IEEE 9th World Forum on Internet of Things (WF-IoT)*, pp. 1–6. IEEE, 2023.

Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, 35:3165–3177, 2022.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Alastair H Moore, Mike Brookes, and Patrick A Naylor. Room geometry estimation from a single channel acoustic impulse response. In *21st European Signal Processing Conference (EUSIPCO 2013)*, pp. 1–5. IEEE, 2013.

Yaniv Nemcovsky, Matan Jacoby, Alex M Bronstein, and Chaim Baskin. Physical passive patch adversarial attacks on visual odometry systems. In *Proceedings of the Asian Conference on Computer Vision*, pp. 1795–1811, 2022.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Giovanni Pintore, Marco Agus, and Enrico Gobbetti. AtlantaNet: Inferring the 3D indoor layout from a single 360° image beyond the Manhattan World assumption. In *European Conference on Computer Vision (ECCV)*, pp. 432–448. Springer, 2020.

Franco P Preparata and Michael Ian Shamos. Convex hulls: Basic algorithms. In *Computational Geometry: An Introduction*, pp. 95–149. Springer, 1985.

Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE transactions on robotics*, 34(4):1004–1020, 2018.

Yuhui Quan, Shijie Deng, Yixin Chen, and Hui Ji. Deep learning for seeing through window with raindrops. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2463–2471, 2019.

Ali Raza, Samia Chelloug, Mohammed Alatiyyah, Jalal Ahmad, and Jeongmin Park. Multiple pedestrian detection and tracking in night vision surveillance systems. *Computers, Materials, & Continua*, 75(2):3275, 2023.

Reuven Y Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112, 1997.

Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 3634–3642. IEEE, 2020.

Manfred R Schroeder. New method of measuring reverberation time. volume 37, pp. 409–412, 1965.

Gabriele Serussi, Tamir Shor, Tom Hirshberg, Chaim Baskin, and Alex M Bronstein. Active propulsion noise shaping for multi-rotor aircraft localization. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 472–479. IEEE, 2024.

Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Daniela Rus. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 5135–5142. IEEE, 2020.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wei Xu, Yixi Cai, Dongjiao He, Jiarong Lin, and Fu Zhang. Fast-lio2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics*, 38(4):2053–2073, 2022.

Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, pp. 146–151. IEEE, 1997.

Zhijian Yang and Romit Roy Choudhury. Maplearn: Indoor mapping using audio, 2024. URL `https://openreview.net/forum?id=PdwrCm5Msr`.

Inmo Yeon, Iljoo Jeong, Seungchul Lee, and Jung-Woo Choi. Echoscan: Scanning complex room geometries via acoustic echoes. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4768–4782, 2024.

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.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

Ji Zhang, Sanjiv Singh, et al. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and systems*, volume 2, pp. 1–9. Berkeley, CA, 2014.

Jindi Zhang, Yang Lou, Jianping Wang, Kui Wu, Kejie Lu, and Xiaohua Jia. Evaluating adversarial attacks on driving safety in vision-based autonomous vehicles. *IEEE Internet of Things Journal*, 9(5):3443–3456, 2021.

Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. Manhattan room layout reconstruction from a single 360° image: A comparative study of state-of-the-art methods. *International Journal of Computer Vision*, 129:1410–1431, 2021.

A Dataset generation

This appendix complements Section 3.2 by reporting the per-scene acoustic statistics, the per-dataset summary table, and the simulator and back-projection bookkeeping that were deferred from the main text. We do not repeat the corpus, simulator, canonical-frame, or split descriptions given in Section 3.2.

Additional simulator details. For ResPlan and AtlantaLayout the source annotation does not provide a ceiling height, so we sample it uniformly from $\mathcal{U}(2.5, 3.0)\,\mathrm{m}$; for Matterport3D the ceiling height is read directly from the layout annotation. For Matterport3D and AtlantaLayout we generate every annotated scene. For ResPlan, we do not generate every possible source–receiver configuration due to the substantially larger layout collection. The resulting subset nevertheless contains 1.48M simulated RIRs, more than either room-scale corpus individually. Atlanta-world floor corners are back-projected through the equirectangular camera model $(\phi, \theta)=(2\pi u, \pi(v-\frac{1}{2}))$, $d=h_{\mathrm{cam}}/\tan\theta$, with a fixed camera height $h_{\mathrm{cam}}=1.6\,\mathrm{m}$.

Sampling Details. From the consolidated pool of Matterport3D, AtlantaLayout and ResPlan, Across all sources, layouts whose floor polygon has area smaller than $2\,\mathrm{m}^2$ are discarded as too small to support meaningful source–receiver sampling on the $0.25\,\mathrm{m}$ grid described below. Each 2D polygon is extruded into a 3D triangular mesh by triangulating floor and ceiling caps at heights 0 and h, and by adding wall faces along polygon edges. Together with material assignment and simulator parameters, this mesh defines the acoustic scene $\mathcal{S}_s$ used to render RIRs. All surfaces use a single homogeneous material from the SoundSpaces 2.0 material database Chen et al. (2022). We do not vary absorption or scattering across surfaces. Source and receiver positions are sampled on a $0.25\,\mathrm{m}$ grid inset by $0.25\,\mathrm{m}$ from walls, both at height $0.55h$. The receiver is offset by $15\,\mathrm{cm}$ behind the source, so the pair is approximately co-located but not coincident. We simulate RIRs using the geometric acoustic ray tracer of Chen et al. (2022) at $f_s = 16\,\mathrm{kHz}$, with duration $0.5\,\mathrm{s}$ ($L_{\mathrm{rir}} = 8000$ samples), 5,000 indirect rays per bounce, diffraction up to order 10, and transmission disabled.

Splits are defined at the scene level rather than the RIR level, ensuring that no train and test samples share wall geometry.

Dataset statistics. Table 3 reports both the total generated-corpus counts and the scene subsets used in our experiments, together with the geometric scale of the resulting polygons and a battery of standard ISO 3382-1 International Organization for Standardization (2009) room-acoustic descriptors. For Matterport3D and AtlantaLayout, all generated scenes are assigned to the reported train/validation/test splits. For ResPlan, the generated corpus contains 868 scenes, of which 287 are used as the held-out active-mapping test set. To keep the per-scene acoustic statistics computable on the full generated corpus, for each scene we sample at most 80 RIRs deterministically (the seed is hashed from the scene name; scenes contributing fewer than 80 RIRs are summarised over all of their RIRs) and, for each descriptor, report the per-scene median over those samples. The per-dataset rows then report the median and IQR of those per-scene medians across the full generated corpus for each source, rather than only the experimental split subsets. Formal definitions of every acoustic and shape descriptor used in the table are given in the next two paragraphs. Table 4 then summarises the closeness of the three corpora as the pairwise 1-D Wasserstein distance between dataset distributions, computed on *per-RIR* samples for the acoustic descriptors (pooled across all sampled RIRs from all generated scenes of each corpus, i.e. at most 80 RIRs per scene) and on *per-scene* scalars for the polygon-shape descriptors. The reported numbers are normalised by the pooled IQR across the three corpora, so that a value of 1.0 corresponds to a shift of one pooled IQR. Each entry is a single point estimate computed on the full per-corpus sample (no bootstrap), so we read it as a coarse separation indicator rather than a hypothesis test.

Acoustic descriptors (formal definitions). Let $h[n]$, $n = 0, \ldots, N-1$, be the broadband mono RIR sampled at $f_s = 16\,\mathrm{kHz}$ (we use the left binaural channel; the right channel is omitted from the per-scene statistics for compactness). All descriptors below are broadband (no octave-band filtering); per-band variants are not used to produce the table. Let $h^2[n]$ denote the squared instantaneous amplitude and let $n_d = \arg\max_{0\le n<N} |h[n]|$ be the index of the largest absolute sample in h, used as the common direct-arrival anchor for clarity, centre time, and DRR. The Schroeder *energy decay curve* Schroeder (1965) is

$$\mathrm{EDC}(n) \;=\; 10\log_{10}\frac{\sum_{m=n}^{N-1} h^2[m]}{\sum_{m=0}^{N-1} h^2[m]} \quad [\mathrm{dB}].$$

The descriptors are then:

- T_{60} (T30 method): linear regression of $\mathrm{EDC}(n)$ over $[-5\,\mathrm{dB},\ -35\,\mathrm{dB}]$, extrapolated to $-60\,\mathrm{dB}$.
- EDT: linear regression of $\mathrm{EDC}(n)$ over the initial $[0\,\mathrm{dB},\ -10\,\mathrm{dB}]$ interval, extrapolated to $-60\,\mathrm{dB}$.
- Clarity index, for $T \in \{50, 80\}\,\mathrm{ms}$ and $n_T = \lfloor T \cdot f_s/1000 \rfloor$:

$$C_T \;=\; 10\log_{10}\frac{\sum_{n\le n_d+n_T} h^2[n]}{\sum_{n>n_d+n_T} h^2[n]} \quad [\mathrm{dB}].$$

  The early/late split is anchored at the direct peak n_d, not at sample 0.
- Centre time, the first temporal moment of h^2 measured from the direct arrival:

$$T_s \;=\; \frac{1}{f_s}\,\frac{\sum_{n=0}^{N-1}(n-n_d)\,h^2[n]}{\sum_{n=0}^{N-1} h^2[n]} \quad [\mathrm{s}].$$

Reported in milliseconds. T_s is signed in principle, but positive in practice for our RIRs because the post-peak tail dominates the energy.

- Direct-to-reverberant ratio, with a symmetric $\pm 2.5\,\mathrm{ms}$ direct window of half-width $n_w = \lfloor 2.5 \cdot f_s/1000 \rceil$ around the peak:

$$\mathrm{DRR} \;=\; 10\log_{10}\frac{\sum_{|n-n_d|\le n_w} h^2[n]}{\sum_{|n-n_d|>n_w} h^2[n]} \quad [\mathrm{dB}].$$

We anchor on n_d rather than the geometric $t_d = d/c$ so that the same definition applies to RIRs whose source–receiver distance is unknown at analysis time. The numbers are mildly sensitive to this window width: at $\pm 1.0\,\mathrm{ms}$ DRR drops by roughly 2–3 dB on all three corpora, while at $\pm 5.0\,\mathrm{ms}$ it rises by a comparable amount; the relative ordering across the three corpora is preserved.

Polygon-shape descriptors (formal definitions). For a closed polygon P with vertices $v_1, \ldots, v_V \in \mathbb{R}^2$, area $A(P)$, perimeter length $L(P) = \sum_{i=1}^{V} \|v_{i+1} - v_i\|$ (indices mod V), and convex hull $\mathrm{conv}(P)$, the table reports:

- Area A (m^2), perimeter L (m), and vertex count V.
- Compactness $4\pi A/L^2$ (Polsby–Popper / isoperimetric ratio): 1.0 for a disk, ≈ 0.785 for a square, smaller for elongated or articulated shapes. We list it for completeness but note that it conflates several distinct shape effects — a long thin corridor and a large multi-room apartment can produce similar values — so when comparing dataset shape distributions we rely more on area, perimeter, and vertex count.
- Convexity ratio $A(P)/A(\mathrm{conv}(P)) \in (0, 1]$ (sometimes called *solidity*): 1 when P is convex; smaller when the polygon has indentations or non-trivial articulations.
- Mean wall length L/V (m): the mean of polygon edge lengths.
- Fraction of non-right interior angles: each interior angle is computed from consecutive edge directions, and is counted as *non-right* when its deviation from the nearest of 90° and 270° exceeds 10° (the 10° tolerance absorbs the small angular noise introduced by the floor-polygon extraction). This descriptor is the cleanest separator between Manhattan corpora (Matterport3D, ResPlan: median 0%) and the panorama-derived Atlanta corpus (median 37%).

The picture that emerges is consistent across descriptors. The broadband reverberation *rate* (T_{60}) is essentially identical across corpora: per-scene medians fall within a 7 ms band (981–986 ms), and the pairwise Wasserstein distance over the per-RIR T_{60} distribution stays below 0.5 pooled IQR even between ResPlan and Matterport3D. The fact that all three corpora produce T_{60} values clustered near 1 s is a direct consequence of the simulator setup of Section 3.2: SoundSpaces 2.0 is run with a single homogeneous interior material across all walls, ceiling, and floor, and ceiling heights are tightly distributed ($\mathcal{U}(2.5, 3.0)$ m for ResPlan and AtlantaLayout, median 3.0 m for Matterport3D), so the absorption per reflection and the geometric mean free path are nearly fixed across the corpus. The remaining acoustic descriptors all set ResPlan apart from the two single-room corpora by $\sim$0.84–1.5 pooled IQR (Tab. 4, MP3D–ResPlan and Atlanta–ResPlan rows), while the in-distribution Matterport3D–AtlantaLayout pair stays at $\sim$0.2–0.7 pooled IQR on the same descriptors — with EDT being the most discriminating intra-cluster descriptor (0.69 IQR) and the four energy-ratio descriptors (C_{50}, C_{80}, T_s, DRR) all near 0.2 IQR. This is the geometric, not physical, signature of ResPlan: with median area $109\,\mathrm{m}^2$ versus $\sim 20\,\mathrm{m}^2$ on the single-room corpora, the time-to-first-reflection in an apartment is much longer at fixed source–receiver offset, which produces a faster initial decay (shorter EDT, smaller T_s) and a higher direct-to-reverberant ratio without changing the late-tail decay rate. We also note that C_{50} and C_{80} carry essentially the same information up to a constant offset (early/late energy ratios at 50 vs. 80 ms) and we report both only because they are the standard ISO 3382-1 pair — the closeness rows for C_{50} and C_{80} in Table 4 are nearly identical, as expected.

The polygon-shape rows of Table 4 confirm the same story from the geometry side: pairwise distances on area, perimeter, and vertex count separate ResPlan from the two single-room corpora by 1.0–1.4 pooled IQR, while AtlantaLayout is uniquely identified by its non-Manhattan corners (median 37% of interior angles deviate from 90° by more than 10°, versus 0% on the other two corpora). Compactness $4\pi A/L^2$ is the one shape descriptor that does *not* cleanly separate ResPlan from the others (closeness 0.4–0.8 pooled IQR), simply because compactness folds together several distinct shape effects, as noted above. The three corpora therefore differ in *shape* and *scale* but share the same simulator physics from SoundSpaces 2.0 Chen et al. (2022) — the statement we relied on in Section 3.2 to justify transferring a model trained on Matterport3D and AtlantaLayout to the held-out ResPlan test set. Figure 4 shows representative polygons grouped by topology family (***Rectangular***, ***Corridors / elongated***, ***Non-Manhattan***, ***Multi-room / articulated***) drawn from all three corpora to make explicit which families each corpus contributes to. Figure 5 then places every individual scene in the (area, vertex-count) plane. Finally, Figure 6 freezes the source position in three rooms of distinct topology and sweeps the microphone along the longest axis of each room, recovering the textbook geometric-acoustics behaviour — a direct-path

arrival time growing linearly with source–receiver distance and a room-dependent reflection structure — which is a strong sanity check that our pipeline produces physically consistent RIRs.

Table 3: Per-dataset summary, with descriptors arranged as rows for direct visual comparison across the three corpora. Scenes and Samples (one RIR per source/microphone pair) report totals over the generated corpus. Experimental scenes by split reports the scene subsets used for train/validation/test experiments. For Matterport3D and AtlantaLayout, these splits cover all generated scenes, whereas 287 of the 868 generated ResPlan scenes are used as the held-out active-mapping test set. Median RIRs / scene reports the distribution over the full generated corpus. All other rows report the per-scene median followed by the per-scene interquartile range (q25, q75) over scenes. The underlying per-scene medians for acoustic descriptors are themselves computed on at most 80 deterministically sampled RIRs per scene. Acoustic descriptors are broadband and follow ISO 3382-1 International Organization for Standardization (2009), with formal definitions given in the corresponding paragraph above. Matterport3D and AtlantaLayout supply the training and matched-distribution test scenes. ResPlan is held out entirely from training, with 287 generated scenes used as the domain-shift active-mapping test set.

	Matterport3D Chang et al. (2017), Zou et al. (2021)	AtlantaLayout Pintore et al. (2020)	*Held-out:* ResI Abouagour & Garyfallidis (20
Counts (totals over the dataset)			
Scenes	2,263	120	
Samples (RIRs)	895,515	33,745	1,476,
Experimental scenes by split (train/val/test)	1629/187/447	92/14/14	0/0/
Median RIRs / scene	253 (120, 502)	202 (123, 374)	1,513 (1,223, 1,9
Polygon geometry (per-scene median, (q_{25}, q_{75}))			
Area (m^2)	22 (12, 40)	17 (11, 30)	109 (90, 1
Max side (m)	5.9 (4.3, 8.6)	5.7 (4.5, 7.5)	13.5 (12.2, 16
Perimeter (m)	20.5 (15.1, 29.5)	16.8 (14.5, 21.9)	50.0 (44.0, 58
Vertex count	4 (4, 8)	6 (4, 10)	12 (10,
Compactness $4\pi A/L^2$	0.41 (0.30, 0.55)	0.41 (0.27, 0.55)	0.45 (0.36, 0.
Convexity ratio	0.99 (0.95, 1.00)	0.94 (0.84, 1.00)	0.86 (0.77, 0.
Mean wall length (m)	3.8 (2.6, 5.0)	2.7 (1.7, 4.6)	3.8 (3.1, 4
Frac. non-right angles ($> 10°$)	0.00 (0.00, 0.00)	0.37 (0.00, 0.85)	0.00 (0.00, 0.
Ceiling height (m)	3.00 (2.73, 3.31)	2.76 (2.63, 2.89)	2.77 (2.63, 2.
Acoustic descriptors (ISO 3382-1, broadband, per-scene median, (q_{25}, q_{75}))			
T_{60} (ms)	986 (872, 1053)	986 (897, 1025)	981 (951, 10
EDT (ms)	816 (663, 918)	864 (792, 949)	340 (289, 4
C_{50} (dB)	6.6 (5.7, 8.1)	5.8 (5.2, 6.8)	11.1 (10.5, 1
C_{80} (dB)	8.7 (8.0, 9.9)	8.0 (7.5, 8.7)	13.0 (12.4, 13
T_s (ms)	27.7 (22.2, 31.9)	31.0 (27.3, 34.2)	12.3 (10.8, 13
DRR (dB)	+1.1 (−0.6, +2.8)	+0.2 (−0.8, +1.7)	+5.1 (+4.5, +5

Table 4: Pairwise dataset closeness, expressed as the 1-D Wasserstein distance between dataset distributions, normalised by the pooled interquartile range across the three corpora. A value of 1.0 corresponds to a shift of one pooled IQR, values close to 0 indicate distributions that are essentially interchangeable. Acoustic distances are computed *per RIR*, pooling at most 80 sampled RIRs from each scene, shape distances are computed *per scene* (one scalar per polygon). Each entry is a single point estimate on the full per-corpus sample (no bootstrap). Matterport3D–AtlantaLayout serves as the in-distribution lower-bound reference. ResPlan is held out. The broadband reverberation rate (T_{60}) is shared across all three corpora, while every descriptor that depends on early/late energy partitioning (EDT, C_{50}, C_{80}, T_s, DRR) and all shape descriptors set ResPlan apart, reflecting its larger envelope size rather than a different reverberation regime.

Descriptor	MP3D–Atlanta	MP3D–ResPlan	Atlanta–ResPlan
Acoustic distributions (per RIR, pooled across ≤ 80 RIRs/scene)			
T_{60} (ms)	0.31	0.50	0.21
EDT (ms)	0.69	1.50	1.11
C_{50} (dB)	0.21	0.94	1.15
C_{80} (dB)	0.22	0.97	1.18
T_s (ms)	0.24	0.87	1.11
DRR (dB)	0.23	0.84	1.07
Polygon shape distributions (per scene, one scalar/polygon)			
Area (m^2)	0.13	1.27	1.39
Perimeter (m)	0.18	1.12	1.29
Compactness $4\pi A/L^2$	0.41	0.44	0.83
Convexity ratio	0.27	0.52	0.75
Vertex count	0.52	1.41	0.95
Mean wall length (m)	0.48	0.29	0.51

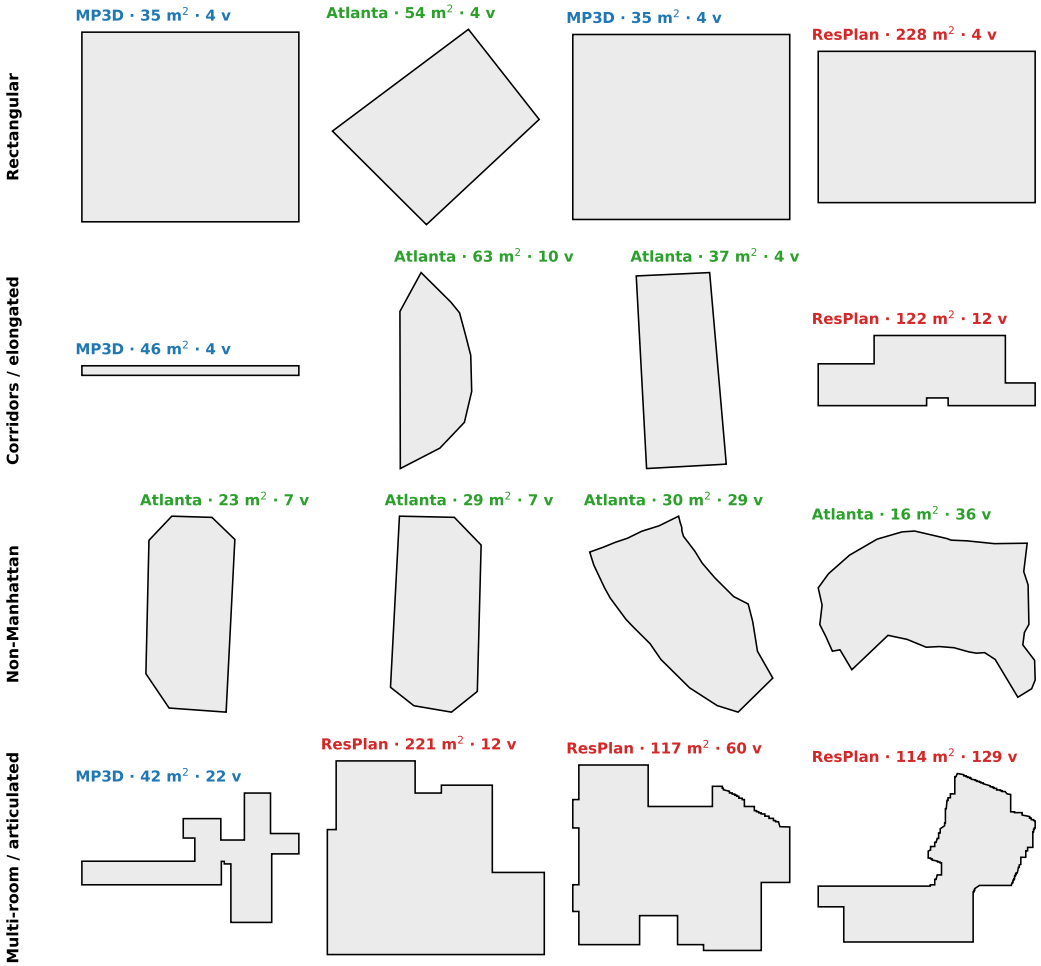


Figure 4: Representative scene polygons grouped by topology. Rows (top to bottom) are: rectangular rooms; corridors and other elongated rooms (aspect ratio ≥ 2); non-Manhattan polygons (single rooms whose corners are *not* aligned to a Manhattan grid); and multi-room or otherwise highly articulated polygons. Within each row the picks are drawn from all three source corpora — the small in-panel tag identifies the corpus (MP3D, Atlanta, ResPlan) along with the polygon area and vertex count — to make explicit that every topology family appears in more than one corpus, and that the corpora differ in the *frequency* with which each family occurs (e.g. Matterport3D dominates the rectangular and corridor rows, AtlantaLayout the non-Manhattan row, and ResPlan the multi-room row), not in exclusive ownership of any one family. The polygon outline is the panel border, drawn in the canonical frame (origin at the top-left, y-axis pointing down, distances in metres).

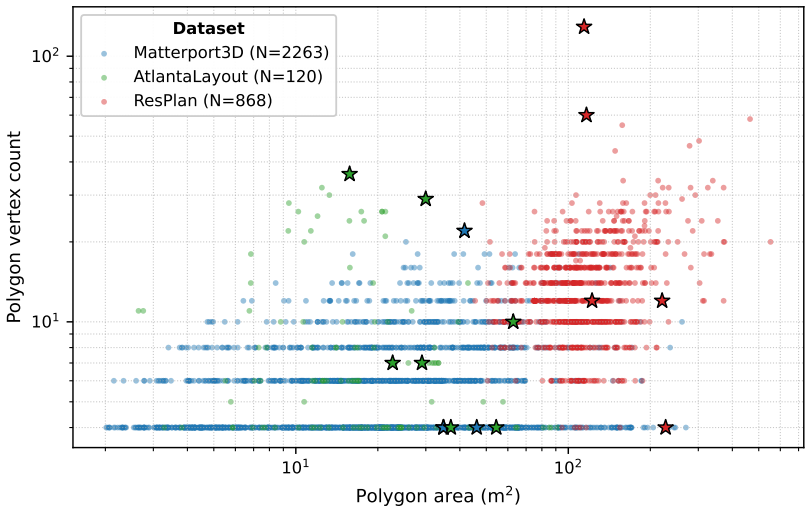


Figure 5: Per-scene polygon area against the polygon vertex count (both axes log-scaled), one dot per scene, coloured by source corpus. Star markers indicate the sixteen curated scenes shown in Figure 4. Matterport3D and AtlantaLayout share the same single-room scale ($\sim$10–40 m^2) and a near-flat bottom edge of 4-vertex rectangles, but AtlantaLayout extends into much higher vertex counts at small areas because of its panorama-derived non-Manhattan corners; ResPlan occupies an almost disjoint regime an order of magnitude larger in area and roughly two-to-five-fold higher in vertex count, with a long upper tail of highly-articulated apartments (≥ 50 vertices, area $> 100\,\mathrm{m}^2$). This plot is the raw counterpart of the area/perimeter/vertex Wasserstein rows of Table 4.

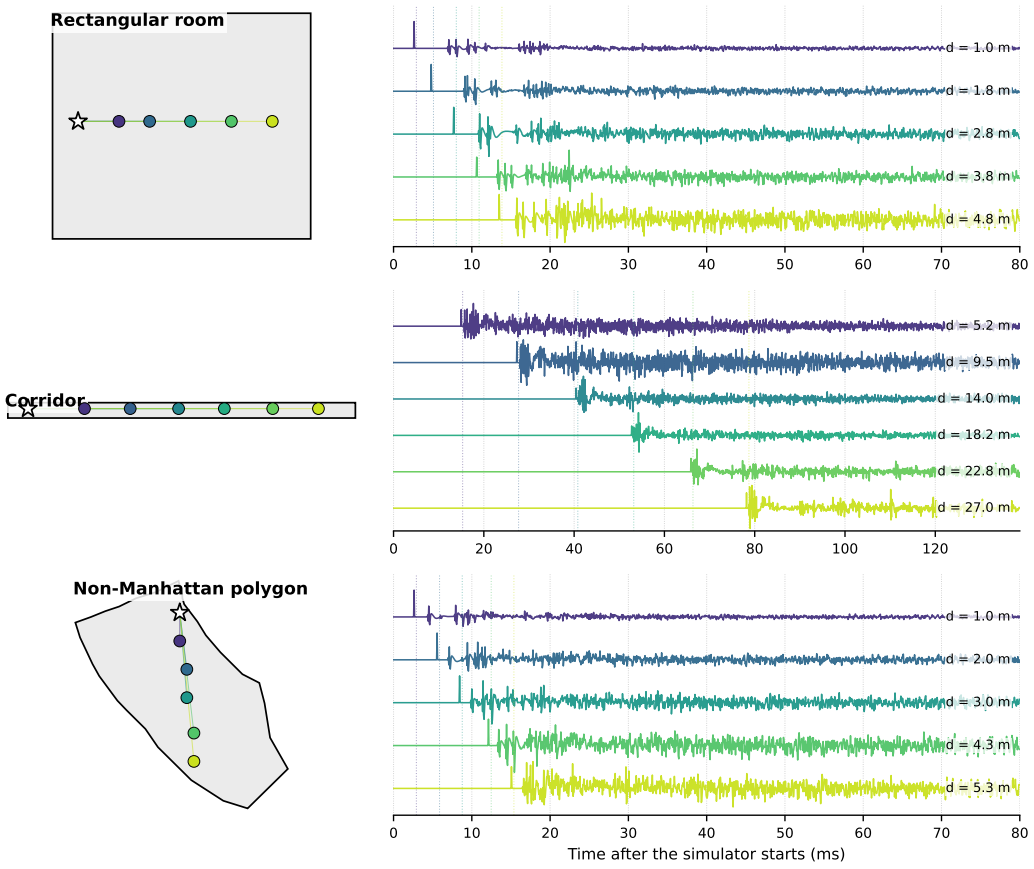


Figure 6: Pipeline sanity check: simulator-generated RIRs in three rooms of contrasting topology. Rows (top to bottom): a rectangular Matterport3D room, an extreme corridor (Matterport3D, $32 \times 1.4\,\mathrm{m}$), and a non-Manhattan AtlantaLayout polygon. In each row the source position is held *fixed* (white star) and five-to-six microphones (coloured circles) are swept along the longest axis of the room; the right column shows the corresponding RIRs, normalised to their own peak and stacked by microphone-from-source distance, with the geometric direct-path arrival d/c marked as a faint vertical line for each trace. Three textbook predictions are recovered: the direct-path peak arrives at d/c (the visible peak in every trace lines up with the predicted vertical line), the peak time grows linearly with distance ($\approx 2.9\,\mathrm{ms/m}$), and the late-tail reflection structure is room-dependent — the corridor produces a long flutter-like tail of widely-spaced reflections from the two end walls, the rectangular room a dense early-reflection cluster, and the non-Manhattan polygon a more diffuse pattern with no clean repetition. This is the simplest end-to-end check that the polygon-extrusion-plus-RLR-Audio pipeline produces physically consistent RIRs from the polygon source alone.

B NOTATION GLOSSARY

To facilitate reading, Table 5 summarizes the primary mathematical notation used throughout the problem formulation and method description.

Table 5: Glossary of mathematical notation.

Symbol	Description
Spaces, Sequences & Environment	
$\mathcal{S}$	The physical acoustic scene (including 3D geometry and materials)
$\mathbb{A}$	The continuous motion action space
a_t	The motion action executed at time step t
p_t	The global 2D pose of the agent at time step t
$\tilde{p}_i$	The relative 2D pose with respect to the start of the current context window
h_t	The interaction history up to time t, defined as $(x_{1:t}, p_{1:t}, a_{1:t-1})$
T_{ep}	The fixed budget of environment steps per episode
$A^{(n)}_{t:t+H_{\text{plan}}-1}$	The n-th candidate action sequence sampled by the planner
H_{plan}	The planning rollout horizon (number of steps)
Acoustic Observations & Latents	
x_t	The monaural room impulse response (RIR) observed at time t
L_{rir}	The number of samples in the raw RIR waveform
u_i	The blind acoustic embedding of x_i
c_i	The Fourier feature embedding of the relative pose $\tilde{p}_i$
S_t	The trajectory-level latent state (world model belief tokens) at time t
$\hat{u}_{t+k}$	The autoregressively predicted future blind acoustic latent
K	The current sliding context window length ($K \leq K_{\max}$)
D	The dimensional size of the latent embeddings
Maps & Spatial Representations	
$M^\star$	The ground-truth global binary occupancy projection of the scene
$m^{\ell\star}_q$	The ground-truth ego-centric local occupancy crop centered at query pose p_q
L_t	The global log-odds spatial belief map at time t
P_t	The global occupancy probability map at time t, where $P_t = \sigma(L_t)$
$\hat{m}^{\ell}_q$	The predicted ego-centric local occupancy map at query pose p_q
$\tilde{L}^{(n)}_{t+H_{\text{plan}}}$	The imagined temporary global log-odds belief after rolling out candidate n
$\mathcal{W}_{p_t}$	The spatial placement operator that projects local maps into the global frame
$H_g \times W_g$	The spatial resolution (height and width) of the global map
$H_\ell \times W_\ell$	The spatial resolution (height and width) of the local ego-centric map
Neural Network Modules	
F_θ	The complete learned inverse acoustic estimator (Acoustic World Model)
e_θ	The blind acoustic Transformer encoder
ρ_θ	The MLP mapping Fourier features to pose embeddings
ϕ_θ	The causal Transformer dynamics model integrating observations and poses
d_θ	The local occupancy decoder
g_θ	The autoregressive future latent predictor
$e_{\bar{\theta}}$	The exponential-moving-average (EMA) copy of the blind acoustic encoder

C Implementation Details and Hyperparameters

This appendix provides the implementation details required to reproduce the local reconstruction and active mapping experiments. Unless otherwise stated, all learned models are trained using the same train/validation/test scene splits, optimization budget, and evaluation protocol. The acoustic world model is trained with distributed data parallelism (DDP) on three NVIDIA L40 GPUs. Planner baselines are evaluated using the same trained acoustic model and global mapping backend unless otherwise specified.

C.1 Training Data Sampling and Target Construction

For local acoustic-to-geometry training, each scene is sampled multiple times across epochs to expose the model to different trajectories, viewpoints, motion patterns, and partial histories of the same underlying geometry. Training trajectories are generated using a momentum-biased random walk over valid source positions. At each step, candidate neighboring poses within a local radius are scored by alignment with the previous motion direction plus random perturbation, producing smooth but diverse motion patterns. Validation and test scenes are sampled with fixed held-out trajectories.

We use variable-length trajectory augmentation during training. Context lengths are sampled between $K_{\min} = 8$ and $K_{\max} = 10$, while evaluation uses $K = 10$ unless otherwise stated. For each trajectory, losses are applied to all valid prefixes starting from the third observation. Each prefix predicts both a local occupancy map and future latent observations, ensuring that the model learns from partial histories while retaining enough context for implicit 2D acoustic trilateration.

Occupancy targets are generated by rasterizing the room polygon into ego-centric local maps centered at the last pose of each trajectory prefix. Each local target covers a $10\,\mathrm{m} \times 10\,\mathrm{m}$ region at 64×64 resolution. Rather than supervising one-pixel binary walls, we thicken the rasterized boundary and apply a small Gaussian blur, yielding soft wall targets. This improves training stability and mitigates the severe imbalance between occupied wall pixels and free space. We further apply random rotation augmentation so that the model learns geometry from different local coordinate frames rather than memorizing absolute orientation.

Table 6: Dataset sampling and local occupancy target construction.

Hyperparameter	Value
Training samples per scene	10
Trajectory step size	0.5m
Minimum context length $K_{\min}$	8
Maximum context length $K_{\max}$	10
Training prefixes supervised from	Step 3
Local map size	10×10m
Local map resolution	64×64

C.2 Acoustic World Model Architecture

Blind Encoder Implementation. This separates pose-agnostic acoustic feature extraction from motion-conditioned spatial reasoning, which is handled later by the trajectory model.

We transform each RIR into a complex STFT and treat STFT frames as tokens. For each frame, the real and imaginary spectral components are concatenated and linearly projected to the model dimension. Using real and imaginary components retains phase information without requiring wrapped phase angles, while preserving spatial cues induced by path-length differences and multipath interference.

The resulting time-frequency token sequence is augmented with Fourier temporal positional encodings: normalized STFT frame indices are embedded using sinusoidal features at multiple frequencies, following Transformer positional encodings Vaswani et al. (2017) and Fourier feature mappings Tancik et al. (2020). This preserves the ordering of arrivals within the RIR, which is important because direct sound, early reflections, and late reverberation encode different geometric constraints. The sequence is then processed by a ViT-style Transformer encoder Dosovitskiy et al. (2020), and a learned $[\mathrm{CLS}]$ token summarizes the full RIR into u_i.

Local Occupancy Decoder Implementation. The decoder uses learnable spatial query tokens arranged on a 2D grid, similar in spirit to query-based decoders used in dense prediction architectures Carion et al. (2020). Each query corresponds to a local map cell and is augmented with spatial coordinate embeddings. The query tokens cross-attend to the world state S_t, and the resulting features are converted to occupancy logits using a convolutional upsampling head. This decoder anchors the latent state to geometry, preventing the predictive

representation from learning acoustic features that are useful for latent matching but irrelevant for mapping. The motivation for decoding local rather than global geometry is twofold. First, prior work shows acoustic evidence is most reliable and informative near the sensing locationLuo et al. (2022). Distant reflections are weaker, more ambiguous, and harder to assign to specific surfaces. Second, local prediction provides a stable supervised target, while global consistency is handled by the mapping backend that fuses local predictions over time. The local map is not restricted to line-of-sight geometry. It may exploit non-line-of-sight acoustic cues, but only within a spatial region where such cues are informative.

Future acoustic latent prediction. To support planning, the model predicts future acoustic embeddings under candidate motions in latent space. Given the current latent belief tokens S_t and a future pose embedding c_{t+k}, the predictor estimates the acoustic latent that would be observed at that pose:

$$\hat{u}_{t+k} = g_\theta(S_t, c_{t+k}), \qquad \hat{u}_{t+k} \in \mathbb{R}^D. \tag{14}$$

The predictor is implemented as a Transformer decoder whose learned query tokens are conditioned on the target pose embedding c_{t+k} and cross-attend to the current world state S_t. The output is an embedding in the acoustic latent space, rather than a reconstructed waveform.

Predicting raw RIR waveforms is poorly matched to active mapping. RIRs are high-dimensional and phase-sensitive: small pose perturbations may produce large waveform differences that are not directly relevant to geometry. Latent prediction instead focuses the dynamics model on stable acoustic structure useful for downstream occupancy inference, following the motivation of joint-embedding predictive representations LeCun et al. (2022), Assran et al. (2025).

Table 7 summarizes the main architectural hyperparameters of the acoustic world model. The model consists of a blind acoustic encoder, a causal dynamics model, a latent predictor, and a spatial query decoder. The blind acoustic encoder converts each RIR into a latent embedding using an STFT-based ViT-style encoder. The causal dynamics model integrates acoustic embeddings with relative pose embeddings. The spatial decoder maps the resulting world state into a local ego-centric occupancy probability map.

Table 7: Acoustic world model architecture hyperparameters.

Hyperparameter	Value
Raw RIR length	4096
Tokenizer type	STFT
STFT n_{fft}	128
STFT window length	32
STFT hop length	8
STFT pad mode	Reflect
Model dimension D	256
Acoustic encoder layers	4
Acoustic encoder heads	8
Acoustic encoder dropout	0.1
Dynamics model layers	6
Dynamics model heads	8
Dynamics model dropout	0.1
Latent predictor layers	4
Latent predictor heads	8
Latent predictor dropout	0.1
Occupancy decoder layers	4
Occupancy decoder heads	8
Decoder coarse grid size	32×32
Decoder output resolution	64×64

C.3 Training Objective

The model is trained with a joint objective combining local occupancy supervision and future latent prediction:

$$\mathcal{L} = \mathcal{L}_{\text{geo}} + \lambda_{\text{pred}} \mathcal{L}_{\text{pred}}.$$

The geometric loss combines focal loss and Dice loss:

$$\mathcal{L}_{\text{geo}} = \lambda_{\text{focal}} \mathcal{L}_{\text{focal}} + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}}.$$

The focal loss addresses wall/free-space imbalance, while the Dice loss encourages spatial overlap of the predicted wall structure.

The latent prediction loss matches predicted future acoustic embeddings to stop-gradient EMA targets:

$$\mathcal{L}_{\text{pred}} = \|\hat{u}_{t+k} - \text{sg}\left(e_{\bar{\theta}}(x_{t+k})\right)\|_2^2 .$$

Table 8: Loss hyperparameters.

Hyperparameter	Value
λ_{pred}	1.0
λ_{focal}	1.0
λ_{dice}	1.0
Focal loss α	0.8
Focal loss γ	2.0

Optimization.

All learned models are trained with AdamW and cosine learning-rate annealing. Training uses gradient accumulation to increase the effective batch size. Gradients are clipped to a maximum norm of 1.0.

Table 9: Optimization hyperparameters for acoustic world model training.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Batch size per process	4
Gradient accumulation steps	4
Number of epochs	300
Learning-rate schedule	Cosine annealing
Gradient clipping	1.0
GPUs	3× NVIDIA L40
Early-stopping patience	40 epochs

EMA target encoder.

The target acoustic encoder is updated using an exponential moving average:

$$\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta.$$

The EMA decay is scheduled during training using a cosine schedule over optimizer steps. The target encoder provides stop-gradient latent targets for the future prediction branch.

Table 10: EMA target encoder hyperparameters.

Hyperparameter	Value
Base EMA decay τ	0.996
Final EMA decay τ	1.0
EMA schedule	Cosine
EMA update frequency	Every optimizer step

C.4 Active Mapping Evaluation Setup

The active mapping benchmark evaluates closed-loop exploration and global map reconstruction. Each episode begins with an initial random warm-up phase of $K = 3$ observations, after which the selected planner controls the agent. All planners run for a maximum of $T = 100$ steps. The environment uses discrete valid acoustic samples with continuous intended motions snapped to nearby feasible navigation nodes.

Table 11: Active mapping environment and mapper hyperparameters.

Hyperparameter	Value
Maximum episode length T	100 steps
Initial warm-up context	3 observations
Motion step size	0.5m
Global canvas size	40×40m

The trained acoustic world model predicts local maps from recent RIR/pose histories. These predictions are fused into a global Bayesian log-odds map. The active mapping metrics are computed on the final global map and explored region.

Planner hyperparameters.

The World-Model CEM planner optimizes short-horizon candidate trajectories using imagined local occupancy maps produced by the acoustic world model. Candidate paths are sampled from a Gaussian distribution over displacement sequences, filtered for feasibility under the current map belief, scored by map-space information gain, and refined over multiple CEM iterations.

Table 12: World-Model CEM planning hyperparameters.

Hyperparameter	Value
Planning horizon H	5
Number of CEM candidates	50
CEM iterations	4
Elite fraction	0.2
Step size	0.5m
Obstacle threshold	0.85
Minimum rollout span	1.5m
Maximum episode steps	100
Collision penalty	15.0
Log-odds clipping	$[-15, 15]$

D Metric Definitions

This appendix defines the metrics used for both local reconstruction and active mapping. All reported means and standard deviations are computed over held-out test scenes unless otherwise stated.

D.1 Thresholding and Boundary Extraction

Predicted occupancy probability maps are thresholded at $\tau = 0.5$ for threshold-dependent metrics unless otherwise stated:

$$\hat{M}_\tau(i,j) = \mathbf{1}\left[\hat{M}(i,j) > \tau\right].$$

For boundary-based metrics, we skeletonize only the ground-truth wall maps, discarding the artificial wall thickening introduced during target construction for training stability. Predicted maps are thresholded but not skeletonized, so overly thick predicted walls are penalized rather than collapsed to a one-pixel boundary. This yields the predicted occupied point set and ground-truth boundary point set

$$\hat{\mathcal{B}} = \{\hat{b}_1, \ldots, \hat{b}_{N_p}\}, \qquad \mathcal{B} = \{b_1, \ldots, b_{N_g}\}.$$

All distances are converted from pixels to meters using the grid resolution Δ.

D.2 Chamfer Distance

We compute symmetric Chamfer distance between the predicted and ground-truth boundary point sets:

$$\mathrm{CD}(\hat{\mathcal{B}}, \mathcal{B}) = \frac{1}{|\hat{\mathcal{B}}| + |\mathcal{B}|} \left(\sum_{\hat{b} \in \hat{\mathcal{B}}} \min_{b \in \mathcal{B}} \|\hat{b} - b\|_2 + \sum_{b \in \mathcal{B}} \min_{\hat{b} \in \hat{\mathcal{B}}} \|b - \hat{b}\|_2 \right) \Delta.$$

Lower Chamfer distance indicates better geometric alignment.

D.3 ϵ-F1

For tolerance ϵ, a predicted boundary point is counted as correct if it lies within ϵ meters of any ground-truth boundary point:

$$\mathrm{TP}_{\mathrm{pred}}(\epsilon) = \sum_{\hat{b}\in\hat{\mathcal{B}}} \mathbf{1}\left[\min_{b\in\mathcal{B}} \|\hat{b} - b\|_2 \Delta \le \epsilon\right].$$

Precision is

$$P_\epsilon = \frac{\mathrm{TP}_{\mathrm{pred}}(\epsilon)}{|\hat{\mathcal{B}}|}.$$

Recall is computed symmetrically by measuring how many ground-truth boundary points are covered by the prediction:

$$\mathrm{TP}_{\mathrm{gt}}(\epsilon) = \sum_{b\in\mathcal{B}} \mathbf{1}\left[\min_{\hat{b}\in\hat{\mathcal{B}}} \|b - \hat{b}\|_2 \Delta \le \epsilon\right],$$

$$R_\epsilon = \frac{\mathrm{TP}_{\mathrm{gt}}(\epsilon)}{|\mathcal{B}|}.$$

The tolerance-based F1 score is

$$F1_\epsilon = \frac{2P_\epsilon R_\epsilon}{P_\epsilon + R_\epsilon}.$$

D.4 Distance-Field MAE

We additionally report distance-field mean absolute error (DF-MAE). For a binary wall map M, let $D_M(i)$ denote the Euclidean distance from grid cell i to the nearest occupied wall cell, measured in meters. We compute

$$\text{DF-MAE} = \frac{1}{|\Omega|} \sum_{i\in\Omega} |D_{\hat{M}}(i) - D_{M^\star}(i)|,$$

where Ω is the evaluated spatial region.

For local reconstruction, Ω is the local prediction window. For active mapping, Ω is restricted to the explored region so that methods are not penalized for unmapped areas outside their observed support.

D.5 Exploration and safety

Coverage.

For active mapping, coverage measures the fraction of ground-truth wall cells that fall inside the explored region:

$$\text{Coverage} = \frac{|\mathcal{B}^\star \cap \Omega_T|}{|\mathcal{B}^\star|} \cdot 100,$$

where $\mathcal{B}^\star$ denotes the full ground-truth wall set and Ω_T is the explored region after T steps.

Spatial efficiency ratio.

Spatial Efficiency Ratio (SER) measures explored area per meter traveled:

$$\text{SER} = \frac{|\Omega_T|\Delta^2}{\sum_{t=1}^{T-1} \|p_{t+1} - p_t\|_2}.$$

Higher SER indicates that an agent covers more unique area per unit distance.

Crash rate.

Crash rate is reported as collisions per 100 meters:

$$\text{CrashRate} = \frac{\#\text{collisions}}{\max(1, \text{PathLength})} \cdot 100.$$

E LOCAL RECONSTRUCTION BASELINES

This appendix provides full descriptions of the baselines listed in Section 4.1. All baselines are evaluated on the same trajectory windows as our model, projected into the same common evaluator frame (a 64×64 ego-centric grid covering 10×10 m, agent at the last pose, y-axis up, walls represented as soft Gaussian bands of width σ=1.5 px), and scored against the same soft-wall ground truth with the same threshold sweep, Chamfer / DF / ϵ-PR/F1 implementations, and plotting code. All outputs are evaluated in this common frame using fixed conversion rules.

E.1 GEOMETRIC TRAJECTORY BASELINES

The geometric baselines (**Convex hull** and **Alpha shape**) compute a 2D polygon directly from the microphone trajectory and ignore the acoustic signal entirely. They quantify the geometry recoverable from the sampled trajectory alone.

Setup. For each context window of K poses we collect the microphone positions $\{m_i\}_{i=1}^{K} \subset \mathbb{R}^2$ in metric coordinates, agent-anchored at the last step, and pass them through one of the two polygon constructors below. The resulting polygon is rasterised into a soft wall probability map by the same renderer that produces the JEPA ground truth (`dataset_jepa.generate_egocentric_maps`): the polygon outline is drawn into the 64×64 / 10 m ego-centric grid, then converted to a soft band by an inverted distance transform passed through a Gaussian of width σ=1.5 px. Pred and GT are therefore pixel-perfectly aligned, so any difference between the two reflects the geometry, not the rasterisation.

Convex hull. We compute $P_{\text{hull}} = \text{ConvexHull}(\{m_i\}_{i=1}^{K})$ via Shapely's `MultiPoint.convex_hull`. This is the simplest possible motion-only prior: it asserts that the agent is always inside the room and that the room is at least the convex polytope it walked. For non-convex layouts, the convex hull can bridge concavities and include space outside the room.

Alpha shape. The alpha shape relaxes convexity. We Delaunay-triangulate the mic point set and keep only triangles whose circumradius is at most

$$\alpha = \alpha_{\text{mult}} \cdot \text{median}_i \min_{j \neq i} \|m_i - m_j\|_2,$$

with α_{mult}=20. The threshold is therefore adaptive to local trajectory density and produces a possibly non-convex polygon that better tracks corridors and bends. We fall back to the convex hull when the trajectory is too short to triangulate (K<3) or when every triangle is rejected.

E.2 ECHOSCAN BASELINE

Original EchoScan. EchoScan Yeon et al. (2024) predicts a global 2D floorplan and a ceiling-height map directly from a small set of RIRs recorded by a fixed multi-channel Ambisonic microphone array placed in a room. Each RIR is encoded by a 1D residual encoder, the per-shot embeddings are aggregated across the small set of recordings into a single descriptor, and a 2D decoder maps the descriptor to a 1024×1024 filled-floorplan probability map covering a ± 10.24 m canvas at 2 cm/px, with the agent at the canvas centre and the image y-axis pointing down. The model is trained with an MSE plus SoftDice loss on the binary floor mask, plus an additional head for ceiling height. *The original work does not release source code*, so we re-implement the architecture from the paper.

Differences from the original paper. Our adaptation matches EchoScan's losses, canvas, optimizer schedule, and RIR pre-processing, and differs from the original paper in three ways that are forced by our setting:

1. **Single-channel input.** Our datasets contain single-channel omnidirectional RIRs, not B-format Ambisonic recordings, so we drop the per-channel processing and rely on a single 1D residual encoder per shot.
2. **Explicit per-shot pose conditioning.** EchoScan assumes a fixed multi-microphone array: spatial information is encoded *implicitly* by the array geometry, which is shared across recordings. We instead vary source/microphone positions across the trajectory, so each shot is augmented with its (x, y) source position (and optionally the receiver position), embedded by a small MLP, and fused with the RIR embedding before set aggregation. This is the only architectural change and is the minimum modification needed to handle a moving sensor.
3. **Floorplan-only head.** Our task is 2D layout reconstruction, so we drop the height head and keep only the floor branch. The loss therefore retains only the floorplan terms, $\mathcal{L} = \text{MSE} + 0.3 \cdot \text{SoftDice}$, with the height term removed.

We do not change the canvas, the encoder, the decoder, or the set-aggregation pooling.

Architecture. For each of K shots, the RIR is encoded by a 1D residual encoder (base width 64, four residual blocks). The source coordinates are embedded by a 128-d MLP. The two streams are concatenated and projected to a shared 256-d feature. The K shot features are aggregated by a multi-aggregation set pooler that concatenates a mean and a generalised-mean (GeM, p=3) along the shot axis. A 2D decoder upsamples the 512-d aggregated descriptor to a 1024×1024 floorplan probability map.

Training. The EchoScan baseline is trained on the Matterport3D and AtlantaLayout train splits used for the world model (Section 3.2), with the same $K \sim 10$ trajectory windows drawn by the JEPA dataset, so all comparisons in Section 4.1 are at parity in data, supervision budget, and trajectory statistics. Targets are rendered directly from the GT polygon onto the 1024×1024 canvas as binary filled masks (no Gaussian thickening). The loss is paper-aligned,

$$\mathcal{L}_{\mathrm{ES}} \;=\; \mathrm{MSE}(\hat{m}, m^{\star}) \;+\; 0.3 \cdot \mathrm{SoftDice}(\hat{m}, m^{\star}),$$

on the floor branch. We optimise with Adam and a cosine-annealing warm-restart schedule (max LR 10^{-3}, min LR 10^{-5}, first cycle 50 epochs, no restarts within the chosen budget), batch size 16, for 200 epochs. RIR pre-processing follows the paper: the first 10 samples of every RIR are muted to suppress the direct path, and no per-RIR normalisation is applied.

Validation convergence. Hyperparameters were selected on validation data. Table 13 reports training and validation performance on the native filled-floorplan output. Validation IoU increases from 0.79429 at epoch 99 to 0.81969 at epoch 199, indicating diminishing gains late in training. Native floor-mask IoU complements the boundary metrics after conversion to the common evaluation frame.

Table 13: Native-output convergence of the adapted EchoScan baseline.

Epoch	Training loss	Training IoU	Validation loss	Validation IoU
0	0.23832	0.20436	0.17904	0.36631
49	0.08427	0.70077	0.06046	0.74570
99	0.06950	0.75109	0.04845	0.79429
149	0.06335	0.77234	0.04321	0.81340
199	0.05920	0.78640	0.04234	0.81969

Post-processing into the evaluation frame. EchoScan emits a *filled-floorplan* probability map at 1024×1024 over ± 10.24 m, y-axis down, while the common evaluator scores soft *wall-presence* maps at 64×64 over 10×10 m, y-axis up, with walls modelled as Gaussian bands of width σ=1.5 px. To compare both models on the same metric pipeline, we project EchoScan's output into the evaluation frame in six fixed steps applied to every batch:

1. Threshold the floor probability at 0.5 to obtain a 1024×1024 binary filled mask.
2. Centre-crop the central 10×10 m window of that mask.
3. Resize the cropped mask to 64×64 with `cv2.INTER_AREA`, then re-binarise at 0.5.
4. Extract a 1-pixel boundary by running `cv2.findContours` on the filled mask and drawing the contour back with `cv2.drawContours` at thickness 1 - exactly the rasterisation that `dataset_jepa.generate_egocentric_maps` uses for the GT walls.
5. Apply a Euclidean distance transform to the inverted boundary and pass it through $\exp\!\big(-d^2/(2\sigma^2)\big)$ with σ=1.5 px to obtain a soft wall map with the same Gaussian profile as the GT.
6. Flip vertically to convert from y-down (graphics) to y-up (evaluation frame).

The result $\hat{p} \in [0,1]^{64\times 64}$ has identical semantics to the JEPA model's raw output, so the evaluator can apply the same threshold sweep, the same Chamfer / DF / ϵ-PR/F1 computations, and the same plotting code to both models without any further normalisation. The same deterministic conversion is applied to every adapted EchoScan output and introduces no learned parameters. Native-space convergence is reported above to complement the common-frame evaluation.

F MEASUREMENT NOISE

We evaluate the trained acoustic-to-geometry model under additive white Gaussian noise at test time, without retraining. Table 14 is evaluated as a separate robustness sweep and therefore includes its own clean reference. The sweep re-evaluates the same trained checkpoint on a separately instantiated held-out trajectory realization, which leads to small numerical differences from the clean results reported in Table 1. All noise levels are

evaluated under the same robustness-sweep protocol, so robustness is assessed relative to the clean reference reported in the same table.

For each noise condition, threshold-dependent metrics use operating thresholds selected on the corresponding validation split, following the same validation-based evaluation protocol used elsewhere in the paper. Model weights remain fixed throughout the sweep. This experiment tests sensitivity to measurement corruption, one component of the simulation-to-real gap.

At 20 and 30 dB SNR, performance is nearly unchanged. At 10 dB, Chamfer increases from 0.146 to 0.157 m and F1 at 0.30 m decreases from 0.958 to 0.934. Degradation becomes larger at 5 dB and below. The results support robustness to moderate additive noise. Transfer across materials, clutter, sensor responses, and structured interference requires further evaluation.

Table 14: Local reconstruction robustness under additive white Gaussian noise without model retraining. Values are mean ± standard deviation. Error metrics are in meters.

Condition	Chamfer ↓	DF-MAE ↓	F1@0.15 m ↑	F1@0.30 m ↑	F1@0.45 m ↑
Clean	0.146 ± 0.087	0.096 ± 0.092	0.816 ± 0.086	0.958 ± 0.077	0.976 ± 0.062
30 dB	0.146 ± 0.086	0.096 ± 0.092	0.816 ± 0.086	0.958 ± 0.076	0.976 ± 0.062
20 dB	0.145 ± 0.083	0.096 ± 0.088	0.816 ± 0.085	0.959 ± 0.075	0.976 ± 0.060
10 dB	0.157 ± 0.082	0.116 ± 0.090	0.772 ± 0.088	0.934 ± 0.078	0.965 ± 0.060
5 dB	0.296 ± 0.148	0.254 ± 0.138	0.579 ± 0.125	0.752 ± 0.134	0.832 ± 0.123
3 dB	0.418 ± 0.225	0.367 ± 0.187	0.482 ± 0.124	0.646 ± 0.145	0.739 ± 0.144
1 dB	0.542 ± 0.278	0.467 ± 0.227	0.408 ± 0.119	0.561 ± 0.148	0.658 ± 0.155

G Planner Details

This appendix provides the full planner definitions used in the active mapping benchmark. All learned and heuristic planners operate through the same interface. The GT-occupancy oracle additionally accesses ground-truth local geometry for candidate scoring. They receive only the current global map belief, the current pose, and, when applicable, access to the learned acoustic world model.

G.1 Global Log-Odds Fusion

The neural model predicts local occupancy probabilities. These predictions are accumulated into a global map using log-odds fusion. Given a local probability map B_t, we first convert it to log-odds:

$$\ell(B_t) = \log \frac{\mathrm{clip}(B_t, \epsilon, 1-\epsilon)}{1 - \mathrm{clip}(B_t, \epsilon, 1-\epsilon)}.$$

The local log-odds map is placed into the global coordinate frame according to the current pose:

$$L_t = L_{t-1} + \mathcal{W}_{p_t}\left(\ell(B_t)\right),$$

where $\mathcal{W}_{p_t}$ denotes the spatial warp from local ego-centric coordinates to the global map canvas.

The global occupancy probability is

$$P_t = \sigma(L_t).$$

To avoid unbounded confidence, log-odds values are clipped to a fixed range. The explored mask Ω_t is updated wherever local evidence is fused into the global map.

G.2 Reactive baselines

Random Walk.

The Random Walk planner samples a random heading independently at every step:

$$\theta_t \sim \mathrm{Uniform}(0, 2\pi),$$

and proposes

$$p_{t+1} = p_t + s \begin{bmatrix} \cos\theta_t \\ \sin\theta_t \end{bmatrix},$$

Algorithm 1 Closed-loop active acoustic mapping

Require: Acoustic world model $F_\theta = (e_\theta, \rho_\theta, \phi_\theta, d_\theta, g_\theta)$, planner π, initial pose p_1, episode budget T_{ep}, warm-up length K_{warm}

Initialize acoustic history $h_1 \leftarrow \emptyset$
Initialize global log-odds map $L_0 \leftarrow 0$
Observe initial RIR x_1 at pose p_1
Initialize history h_1 with x_1, p_1 and an empty action sequence
for $t = 1, \ldots, K_{\mathrm{warm}} - 1$ **do**
    Sample warm-up action a_t^{warm}
    Execute a_t^{warm} in the environment
    Observe resulting pose p_{t+1}, RIR x_{t+1}, and collision flag
    if no collision **then**
        Append x_{t+1}, p_{t+1} and the executed action to the history
    else
        Keep current pose and resample warm-up action
    end if
end for
for $t = K_{\mathrm{warm}}, \ldots, T_{\mathrm{ep}}$ **do**
    Extract recent RIRs and poses $(x_{t-K+1:t}, p_{t-K+1:t})$ from the history
    Encode RIRs and relative poses into world state

$$S_t = \phi_\theta(e_\theta(x_{t-K+1:t}), c_{t-K+1:t})$$

    Decode current local occupancy map

$$\hat{M}_t^{\mathrm{loc}} = d_\theta(S_t, \tilde{p}_t)$$

    Fuse $\hat{M}_t^{\mathrm{loc}}$ into global log-odds map L_t
    Query planner for next action

$$a_t = \pi(h_t, L_t, p_t, F_\theta)$$

    if a_t is terminal **then**
        break
    end if
    Execute action a_t in the environment
    Observe resulting pose p_{t+1}, RIR x_{t+1}, and collision flag
    if collision occurs **then**
        Insert collision site into L_t as occupied evidence
    end if
    Append x_{t+1}, p_{t+1} and the executed action to the history
end for
return final global map L_T

where s is the step size. This planner does not use the map or model for action selection.

DFS heuristic.

The DFS heuristic maintains a persistent heading and continues moving in that direction until the current map predicts an obstacle. A direction is considered blocked if the ray from the current pose to the proposed next pose intersects any cell with occupancy probability above a threshold:

$$P_t(i, j) > \tau_{\text{obs}}.$$

When blocked, the planner samples a new random direction and repeats this test. If no clear direction is found after a fixed number of attempts, the planner terminates.

G.3 FRONTIER + A*

The Frontier + A* planner implements classical frontier-based exploration Yamauchi (1997). Free space is defined as explored cells with sufficiently low occupancy probability:

$$\mathcal{F}_t = \{(i, j) : P_t(i, j) < 0.5,\ (i, j) \in \Omega_t\}.$$

Unexplored cells are those outside the exploration mask:

$$\mathcal{U}_t = \{(i, j) : (i, j) \notin \Omega_t\}.$$

Frontier cells are unexplored cells adjacent to explored free space:

$$\mathcal{R}_t = \text{Dilate}(\mathcal{F}_t) \cap \mathcal{U}_t.$$

Connected frontier components are ranked by size. For each component, the planner selects the closest frontier pixel to the agent and plans a path using A* Hart et al. (1968). Cells with occupancy probability above τ_{obs} are treated as obstacles. If all frontiers are unreachable, the planner terminates.

Frontier + A* (WM).

The Frontier + A* (WM) planner extends the classical frontier approach by incorporating predictive safety checks using our acoustic world model. Candidate frontier regions $\mathcal{R}_t$ and initial A* paths are extracted from the current spatial belief exactly as in the standard baseline. However, before committing to a path, the planner queries the world model to predict the sequence of future acoustic observations along the intended trajectory. These predicted observations are decoded and recursively fused into a predicted future occupancy belief. If the intended path intersects with predicted obstacles (i.e., cells exceeding τ_{obs} in the predicted future map), the trajectory is deemed unsafe and discarded, prompting the planner to evaluate the path to the next ranked frontier. These predictive checks reduce mean collision rate in Table 2, with reduced coverage under the fixed budget.

G.4 CEM (WM)

Our planner uses the acoustic world model to evaluate candidate future motion sequences. At decision time t, CEM maintains a Gaussian distribution over action sequences:

$$A^{(n)}_{t:t+H-1} \sim \mathcal{N}(\mu, \Sigma),$$

where H is the rollout horizon.

Before sampling CEM candidates, the planner performs the same reachability sanity check used by the Frontier + A* planner: if no frontier is reachable under the current global belief, the episode is terminated. This prevents the planner from optimizing over trajectories when the current map indicates that no unexplored region can be safely reached.

For each CEM iteration, candidate trajectories are generated using rejection sampling. A sampled candidate is rejected if the ray-cast path from the current pose intersects a high-confidence occupied cell in the current belief map, or if the total spatial displacement between the start and end of the candidate trajectory is smaller than a minimum span threshold. We set this minimum span to 1.5m to reject collapsed candidates that would merely dither locally. Sampling continues until the desired number of valid candidates is obtained or until a maximum number of attempts is reached. In our implementation, this limit is set to $20N$, where N is the number of CEM candidates.

If rejection sampling fails to produce enough valid candidates to form the elite set, i.e., fewer than $N \cdot \rho$ candidates are found for elite fraction ρ, the planner falls back to the Frontier + A* policy and takes a step toward

a reachable frontier. This fallback handles cases where the CEM proposal distribution is overly constrained by the current belief, while still preserving safe exploration behavior.

For each remaining valid candidate, the world model predicts future local occupancy maps:

$$\hat{M}^{\mathrm{loc},(n)}_{t+1:t+H}.$$

These imagined local maps are fused into a temporary copy of the current global log-odds map, producing an imagined future belief $\tilde{L}^{(n)}_{t+H}$ and probability map

$$\tilde{P}^{(n)}_{t+H} = \sigma(\tilde{L}^{(n)}_{t+H}).$$

The score is based on map-space entropy reduction:

$$R(A^{(n)}) = \sum_{i,j} \left[\mathcal{H}(P_t(i,j)) - \mathcal{H}(\tilde{P}^{(n)}_{t+H}(i,j)) \right],$$

where

$$\mathcal{H}(p) = -p \log p - (1-p)\log(1-p).$$

During imagined rollout, the planner also performs collision checking in the temporary map. If a candidate is predicted to collide with a high-confidence occupied region, the trajectory is severely penalized in its information-gain score. In implementation, the score also includes a boundary-discovery bonus that upweights information gain near predicted occupied structure.

After scoring, the top elite fraction of candidates is selected:

$$\mathcal{E} = \mathrm{TopK}\left(\{A^{(n)}\}, R \right).$$

The CEM distribution is refit to the elite set:

$$\mu \leftarrow \mathrm{mean}(\mathcal{E}), \qquad \Sigma \leftarrow \mathrm{cov}(\mathcal{E}).$$

After a fixed number of CEM iterations, the planner executes only the first action of the best sequence, observes the real RIR, updates the map, and replans.

CEM (No-WM).

The CEM (No-WM) control retains the CEM optimizer, horizon, rejection sampling constraints, and Frontier + A* fallback of CEM (WM). It scores valid trajectories by proximity to frontiers extracted from the current global belief L_t. Both variants use the same learned local mapper for observed RIRs. The comparison evaluates predictive information-gain and safety scoring against frontier guidance under the matched optimization framework.

G.5 GT-OCCUPANCY ORACLE

The oracle uses the CEM framework and map-space entropy scoring of CEM (WM). At each candidate future pose, it retrieves the ground-truth local occupancy crop from the simulator and fuses that crop into the temporary belief for information-gain and collision evaluation. Executed observations use the shared learned mapping backend. This reference removes errors in predicted future geometry during scoring, including errors from acoustic prediction and occupancy decoding. Finite-horizon search and the learned mapping backend still affect its performance.

Collision feedback.

If the executed action results in a physical collision, the collision location is inserted into the global map as strong occupied evidence. This acts as a tactile correction signal and prevents repeated attempts to cross the same boundary.

H TRAJECTORY LENGTH ABLATION

A single omnidirectional RIR from a single source-receiver pair provides only ambiguous distance evidence about the scene. Echo arrivals constrain possible reflector locations, but they do not uniquely determine wall geometry as many different reflector configurations can produce similar arrival patterns from a single sensing pose. Recovering geometry therefore requires motion. By observing how the acoustic response changes across a

Algorithm 2 World-Model CEM Planner

Require: Current context h_t, global map L_t, current pose p_t, world model F_θ, frontier fallback planner π_{frontier}
Require: Horizon H, candidates N, CEM iterations J, elite fraction ρ
Require: Max sampling attempts M_{attempt}, minimum candidate span d_{span}

Check whether any frontier is reachable under the current belief L_t
if no reachable frontier exists **then**
    return terminal status
end if
Initialize CEM distribution over action sequences $\mathcal{N}(\mu, \Sigma)$
Set $N_{\text{elite}} \leftarrow \lceil \rho N \rceil$
for $j = 1, \ldots, J$ **do**
    Initialize valid candidate set $\mathcal{C} \leftarrow \emptyset$
    Initialize attempt counter $m \leftarrow 0$
    while $|\mathcal{C}| < N$ and $m < M_{\text{attempt}}$ **do**
        Sample candidate action sequence

$$A^{(m)}_{t:t+H-1} \sim \mathcal{N}(\mu, \Sigma)$$

        Convert $A^{(m)}_{t:t+H-1}$ to future poses

$$p^{(m)}_{t+1:t+H}$$

        if candidate path crosses high-confidence occupied cells in L_t **then**
            Reject candidate
        else if candidate spatial span is smaller than d_{span} **then**
            Reject candidate
        else
            Add $A^{(m)}_{t:t+H-1}$ to $\mathcal{C}$
        end if
        $m \leftarrow m + 1$
    end while
    if $|\mathcal{C}| < N_{\text{elite}}$ **then**
        return first action from Frontier + A* fallback π_{frontier}
    end if
    for each candidate $A^{(n)}_{t:t+H-1} \in \mathcal{C}$ **do**
        Roll out future acoustic latents with F_θ along $p^{(n)}_{t+1:t+H}$
        Decode imagined local occupancy maps

$$\hat{M}^{\text{loc},(n)}_{t+1:t+H}$$

        Initialize temporary belief $\tilde{L}^{(n)}_t \leftarrow L_t$
        Fuse imagined maps into temporary belief

$$\tilde{L}^{(n)}_{t+H}$$

        Compute map-space information-gain score

$$R(A^{(n)}) = \sum_{i,j} \left[\mathcal{H}(\sigma(L_t(i,j))) - \mathcal{H}(\sigma(\tilde{L}^{(n)}_{t+H}(i,j))) \right]$$

        if candidate is predicted to collide in the imagined belief **then**
            Apply severe collision penalty to $R(A^{(n)})$
        end if
    end for
    Select elite candidates

$$\mathcal{E} \leftarrow \text{TopK}_\rho(\mathcal{C}, R)$$

    Refit CEM distribution to elite set

$$\mu \leftarrow \text{mean}(\mathcal{E}), \qquad \Sigma \leftarrow \text{cov}(\mathcal{E})$$

end for
Let $A^\star = \arg\max_{A^{(n)} \in \mathcal{C}} R(A^{(n)})$
return first action from $A^\star$

spatial baseline, the model can implicitly triangulate or trilaterate geometric structure from multiple acoustic views.

This ablation quantifies how much trajectory context is required for local acoustic-to-geometry reconstruction. All metrics in this ablation are reported on the held-out RESPLAN dataset.

For each context length

$$K \in \{2, 4, 6, \ldots, 20\},$$

we take the final K observations from the same trajectory and predict the local occupancy map at the final pose. The prediction target is therefore held fixed across all values of K, isolating the effect of historical context length.

For each sliced trajectory, coordinates are re-centered so that the first pose of the slice is the origin:

$$\tilde{p}_i = p_i - p_{t-K+1}.$$

This matches the relative-coordinate convention used during training and avoids introducing artificial out-of-distribution coordinate offsets. Since the model was trained with a maximum context length of $K_{\max} = 10$, evaluations beyond $K = 10$ test whether the model can extrapolate to longer acoustic histories than those seen during training.

Figure 7 reports Chamfer distance, DF-MAE, and $F1_\epsilon$ as a function of context length. Figure 8 provides qualitative examples, where each row corresponds to a fixed trajectory and each column shows the reconstruction obtained from a different number of observations. The rightmost column shows the ground-truth local map.

The results show two regimes. First, performance improves substantially as the context grows from very short histories to the training context length. This confirms the physical intuition that a small motion baseline is insufficient to resolve acoustic ambiguity, while additional observations provide complementary constraints. Second, performance degrades once the context length substantially exceeds the training window. This degradation is expected: the model was optimized for contexts up to $K_{\max} = 10$, and longer sequences change both the token distribution and the attention pattern of the dynamics model. In practice, this motivates the sliding-window design used throughout the paper: the agent maintains a bounded recent acoustic history rather than attending over an arbitrarily long trajectory.

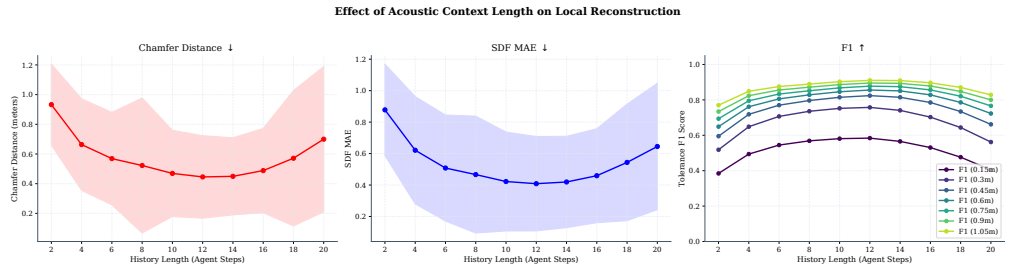


Figure 7: Effect of acoustic context length on local reconstruction. We vary the number of observations K provided to the model and evaluate Chamfer distance, DF-MAE, and $F1_\epsilon$ on held-out RESPLAN trajectories. Performance improves as additional motion-conditioned acoustic observations reduce ambiguity, then degrades when the context length exceeds the training window $K_{\max} = 10$.

I AUTOREGRESSIVE ROLLOUT DRIFT

We evaluate how well the world model can imagine future geometry under autoregressive rollout. Starting from a short warm-up context of two real RIR observations, we unroll the predictor forward for horizons $H = 1, \ldots, 10$. At each imagined step, the predicted latent is fed back into the model to produce the next latent and the corresponding local occupancy map. We compare these imagined maps against ground-truth local occupancy maps at the corresponding future poses.

To separate rollout error from the intrinsic ambiguity of the acoustic signal, we compare autoregressive rollout against a *pure reconstruction* reference, in which all observations are real and the predictor is not used. The pure reconstruction curve therefore reflects the geometry that can be inferred when the model is given the true future acoustic evidence, while the autoregressive curve measures how much quality degrades when future observations must be imagined. As expected, prediction quality degrades with horizon due to compounding error. However, the world model remains capable of producing geometrically meaningful future beliefs over the short horizons relevant for planning.

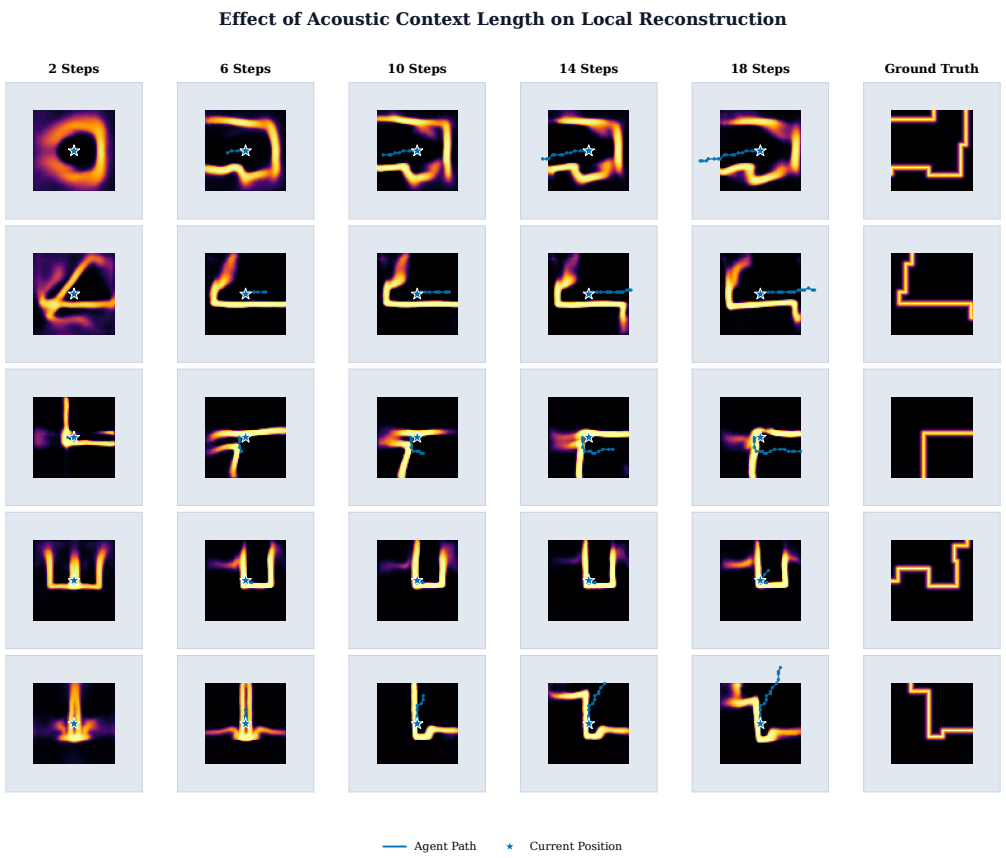


Figure 8: Qualitative trajectory length ablation. Each row corresponds to a fixed trajectory, and each column shows the predicted local occupancy map using a different number of observations. The rightmost column shows the ground-truth local map. Short contexts produce ambiguous or incomplete geometry, while moderate contexts provide enough spatial baseline for sharper reconstruction. Very long contexts can introduce blur or misalignment because they exceed the training context length.

All metrics in this ablation are reported on the held-out RESPLAN dataset.

Figure 9 reports Chamfer distance and $F1_\epsilon$ for $\epsilon \in \{0.15, 0.45\}$ as a function of rollout horizon, comparing autoregressive rollout against pure reconstruction. Together, these results show that although drift accumulates over time, the predictor remains effective enough to imagine useful future world states for short-horizon planning.

Figure 10 shows qualitative examples for a rollout horizon of $H = 5$. In each example, we compare ground truth, pure reconstruction, and world-model rollout. The qualitative comparison shows that the predictor often produces plausible future geometry, but can sometimes inherit or preserve ambiguity already present in the acoustic evidence. This behavior is consistent with the quantitative results and motivates our use of short-horizon receding-horizon planning.

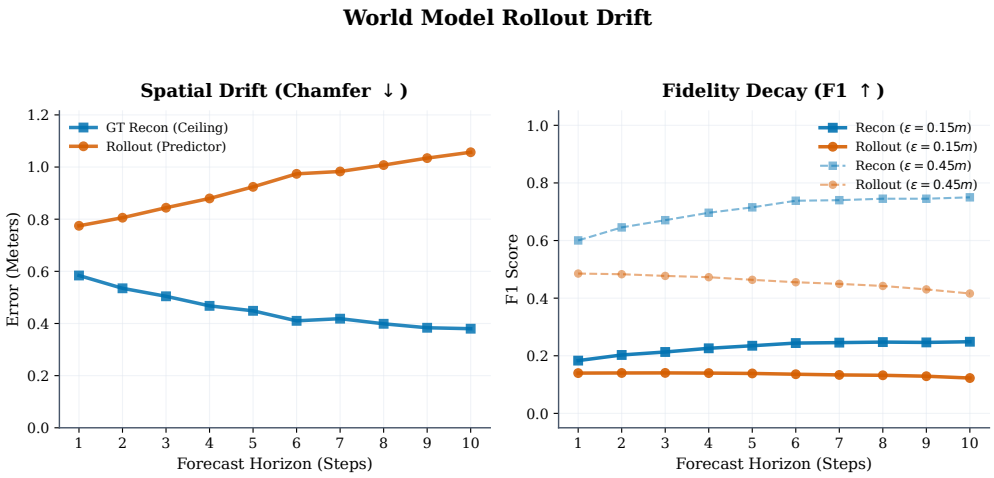


Figure 9: Autoregressive rollout drift as a function of horizon. We compare imagined rollouts against a pure reconstruction reference in which all observations are real. The plot reports Chamfer distance, $F1_{\epsilon=0.15}$, and $F1_{\epsilon=0.45}$ across horizons $H = 1, \ldots, 10$.

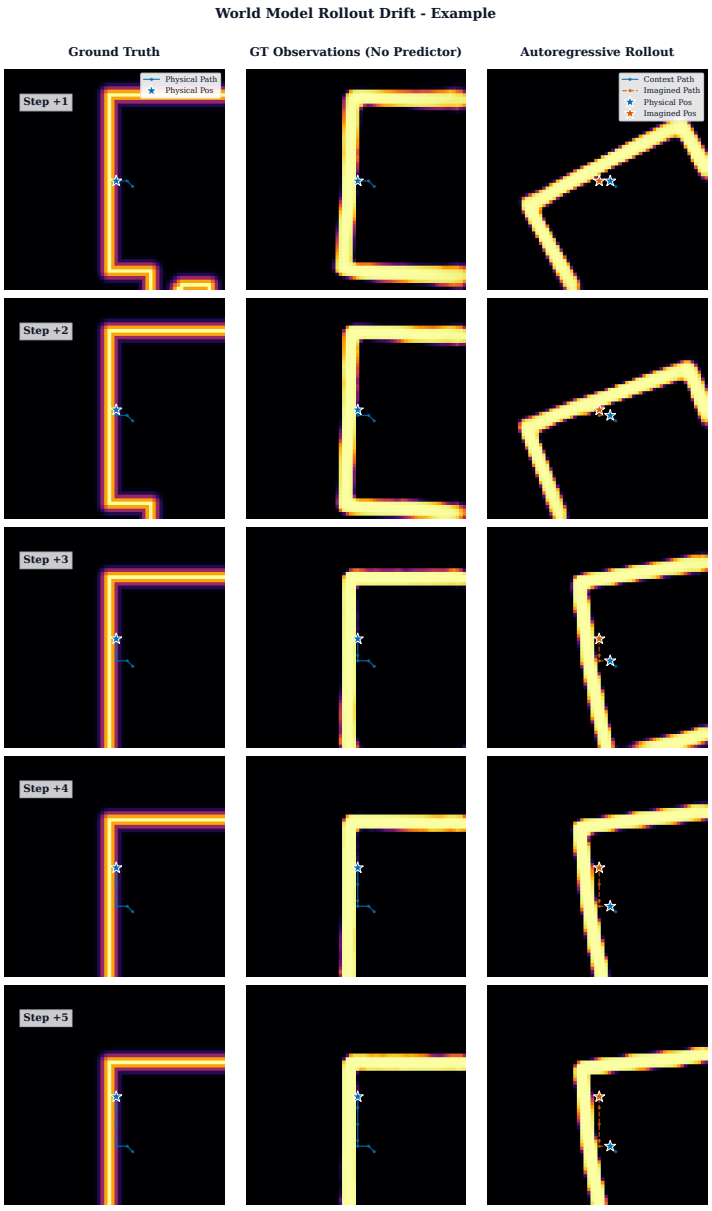


Figure 10: Qualitative comparison of future geometry prediction at rollout horizon $H = 5$. Each rollout starts from two real RIR observations as warm-up. We compare ground truth, pure reconstruction using real future observations, and autoregressive world-model rollout. The predictor often imagines plausible future geometry, but errors can arise when the initial acoustic context remains ambiguous or when the rollout accumulates small pose-alignment errors, leading to local rotations or translations relative to the ground-truth map.

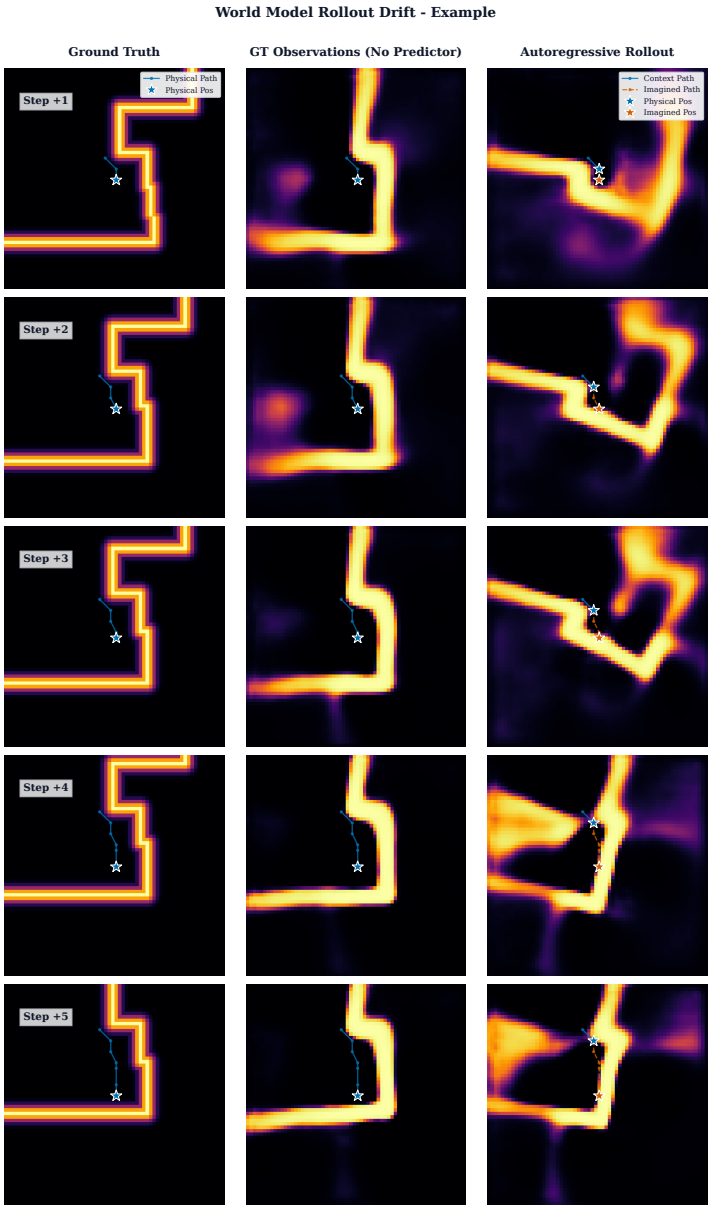


Figure 10: Qualitative comparison of future geometry prediction at rollout horizon $H = 5$ (continued).

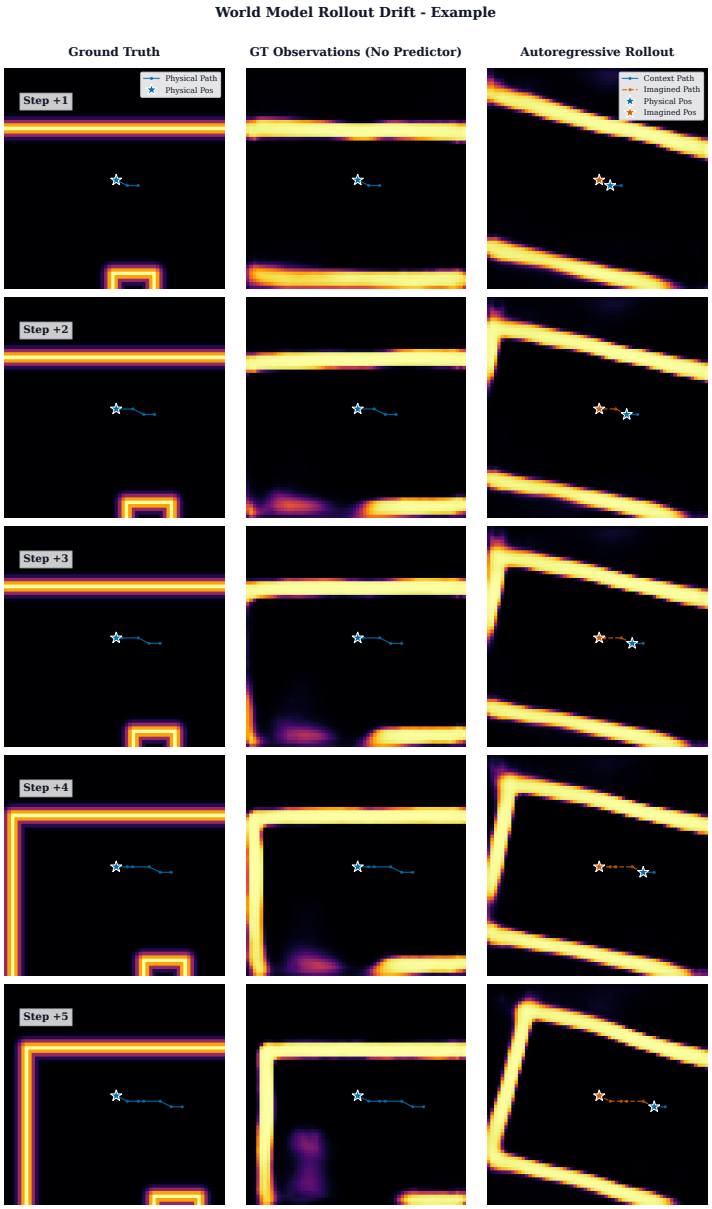


Figure 10: Qualitative comparison of future geometry prediction at rollout horizon $H = 5$ (continued).

J Extended Local Reconstruction Examples

We provide additional qualitative examples of local acoustic-to-geometry reconstruction on held-out scenes. These examples illustrate how the model behaves across different room geometries, including cases where motion-conditioned acoustic history produces sharp local maps and cases where acoustic ambiguity leads to missing or shifted wall estimates.

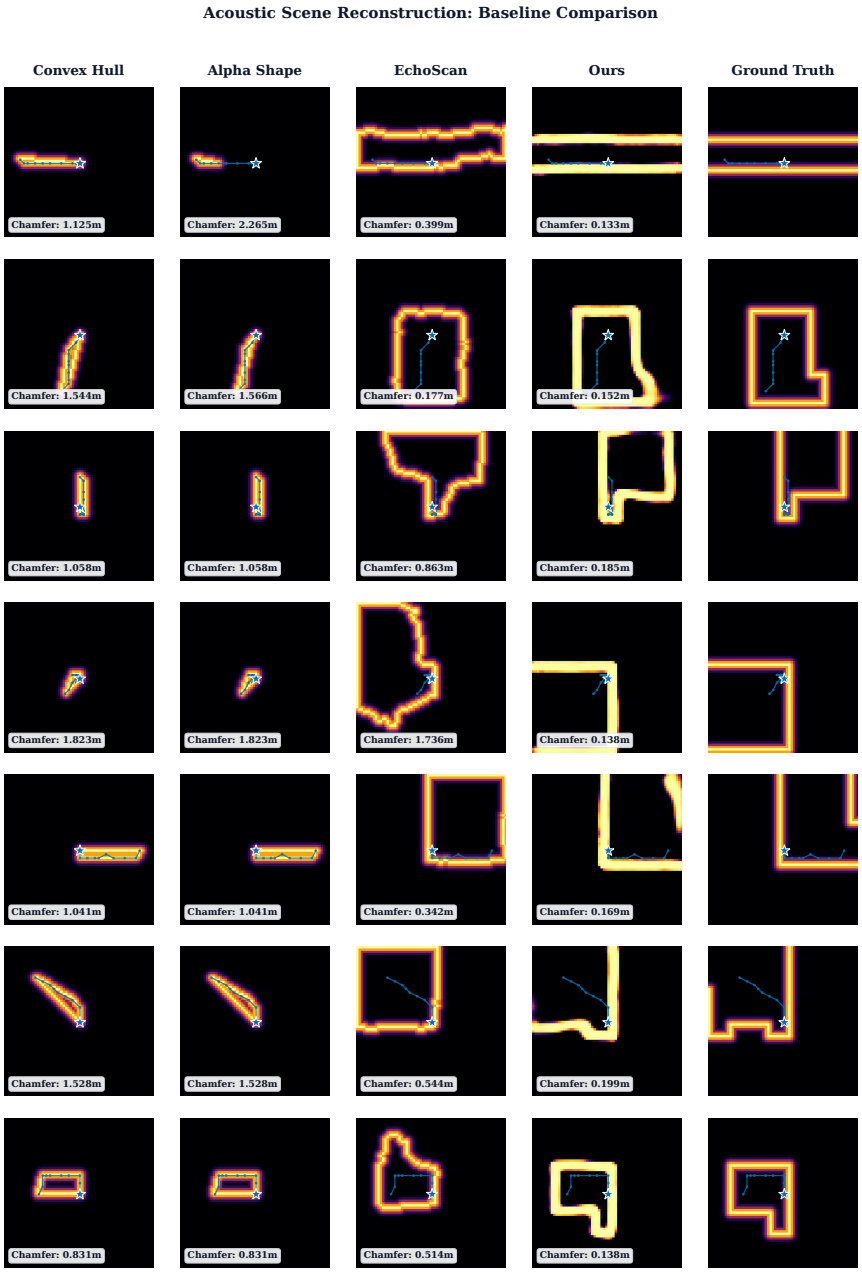


Figure 11: Qualitative local acoustic-to-geometry reconstruction on held-out scenes. Each row shows one example using a trajectory length of $K = 10$ observations. The columns correspond to different baselines and methods, with the rightmost column showing the ground-truth local map. In the reported figure, the baselines include Alpha Shape (with $\alpha = 20$) and EchoScan. Our model produces sharper and more geometrically consistent wall estimates.

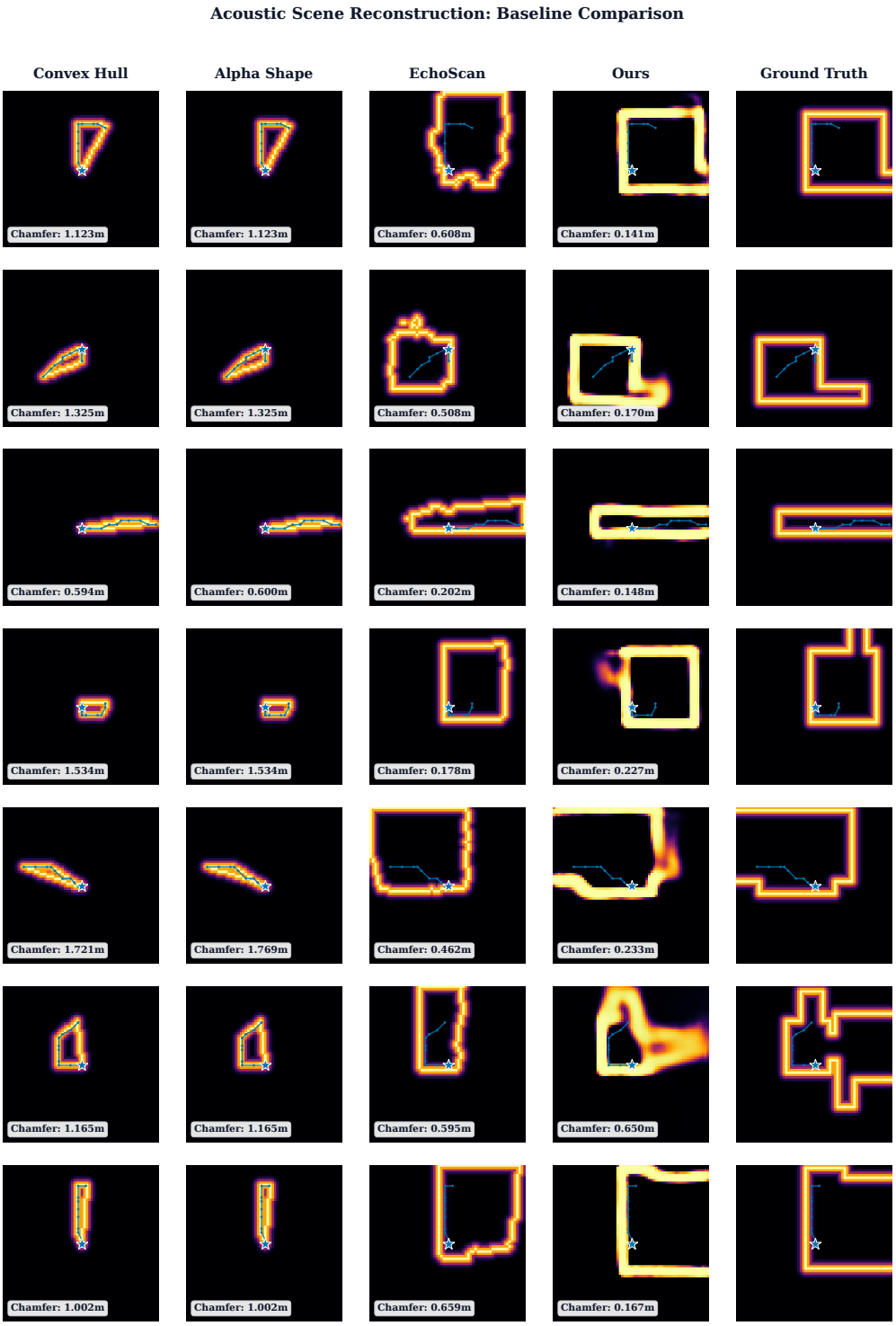


Figure 11: Qualitative local acoustic-to-geometry reconstruction on held-out scenes (continued).

K Extended Active Mapping Results

We provide additional qualitative results for the closed-loop active mapping benchmark. The following subsections separately visualize final mapping behavior across planners and the intermediate CEM rollouts used by the proposed planner.

K.1 Additional Active Mapping Examples

We show additional active mapping episodes across held-out scenes. Each example visualizes the executed trajectory, start and end locations, crash sites, and final predicted global map for the compared planners, highlighting how different action-selection policies affect coverage and reconstruction.

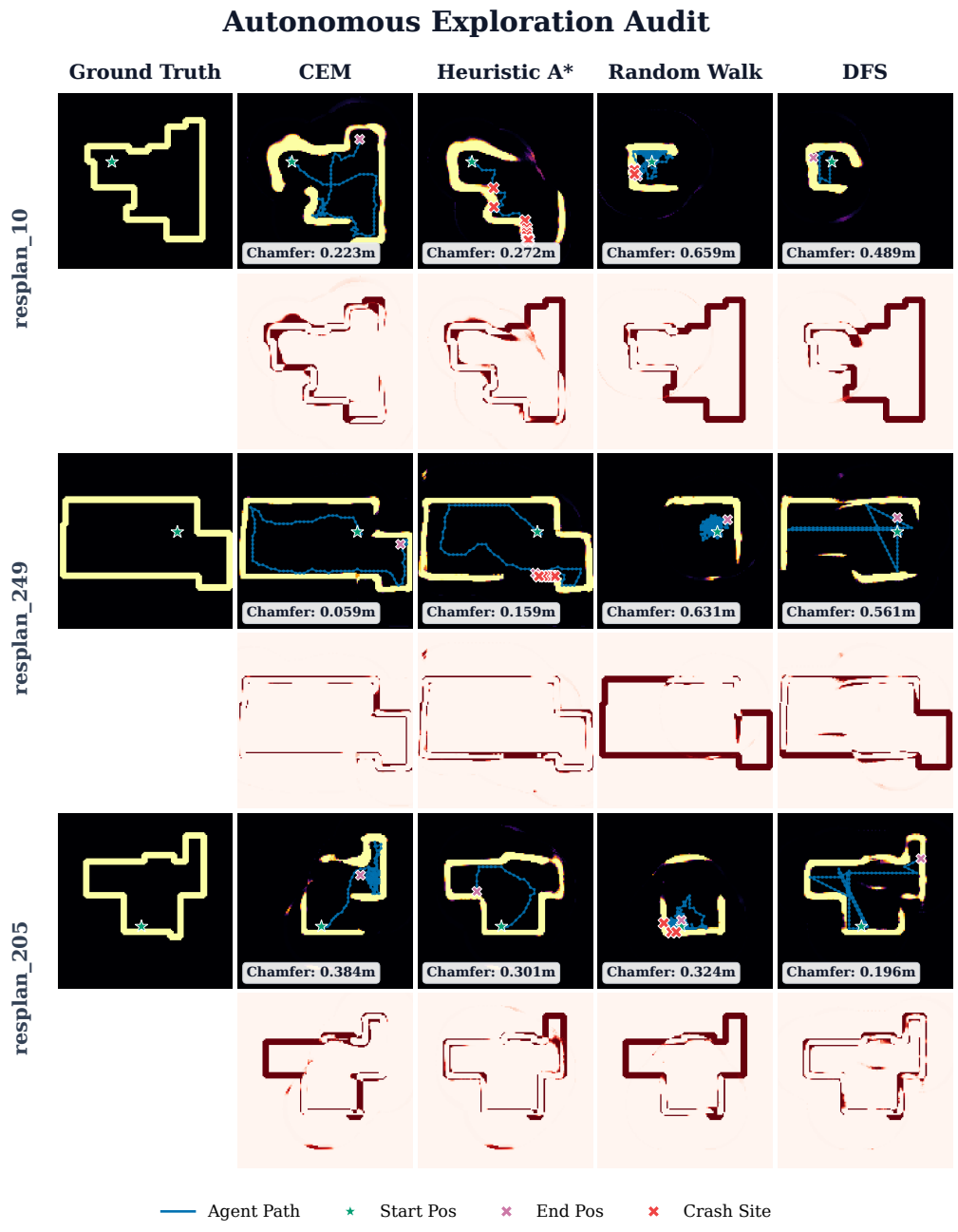


Figure 12: Qualitative active acoustic mapping comparison on held-out scenes. Each row shows one scene, and columns correspond to different planners using the same learned acoustic mapper and global log-odds fusion backend. Panels show the final predicted global map together with the executed trajectory, start location, final location, and crash sites.

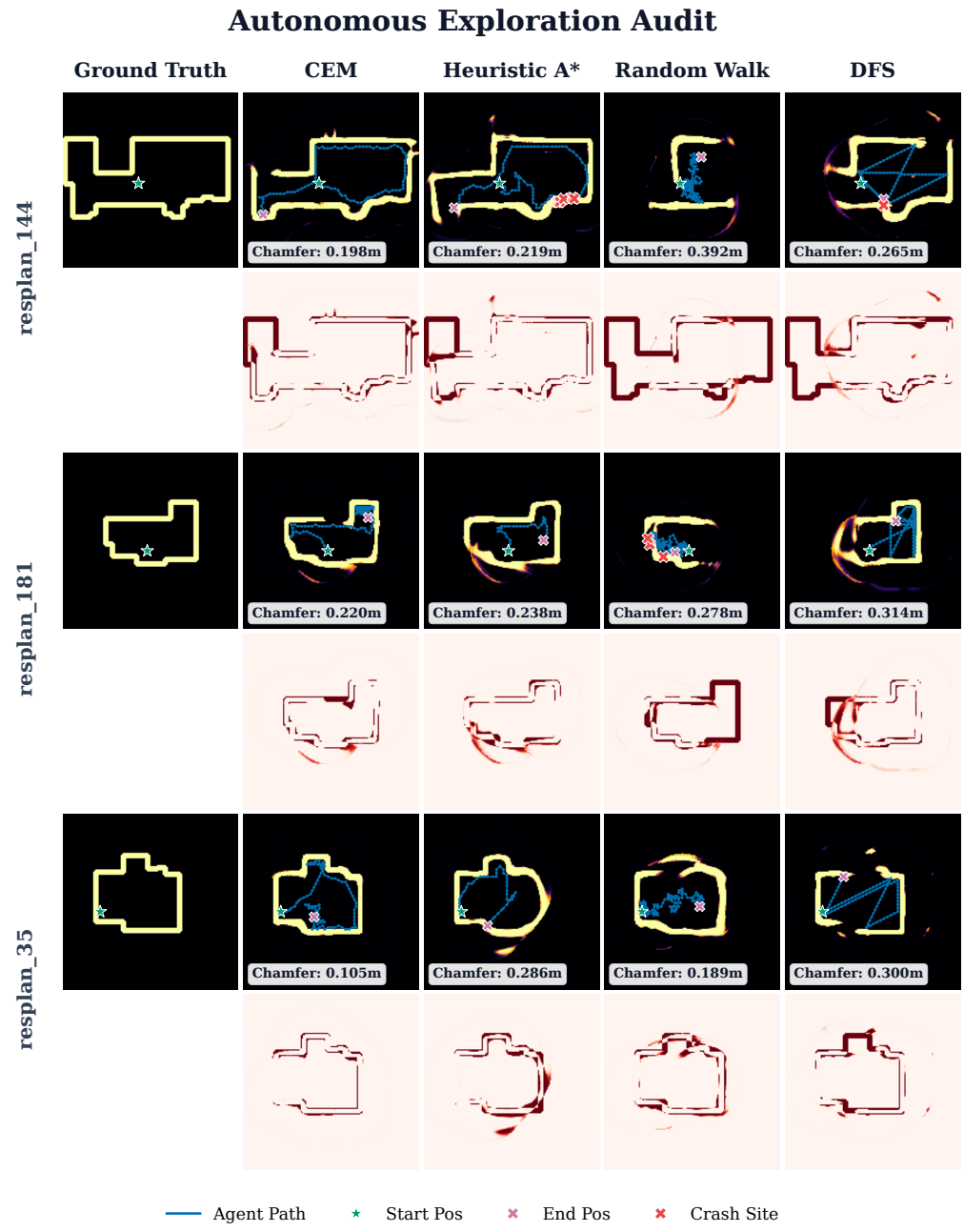


Figure 12: Qualitative closed-loop active acoustic mapping comparison on held-out scenes (continued).

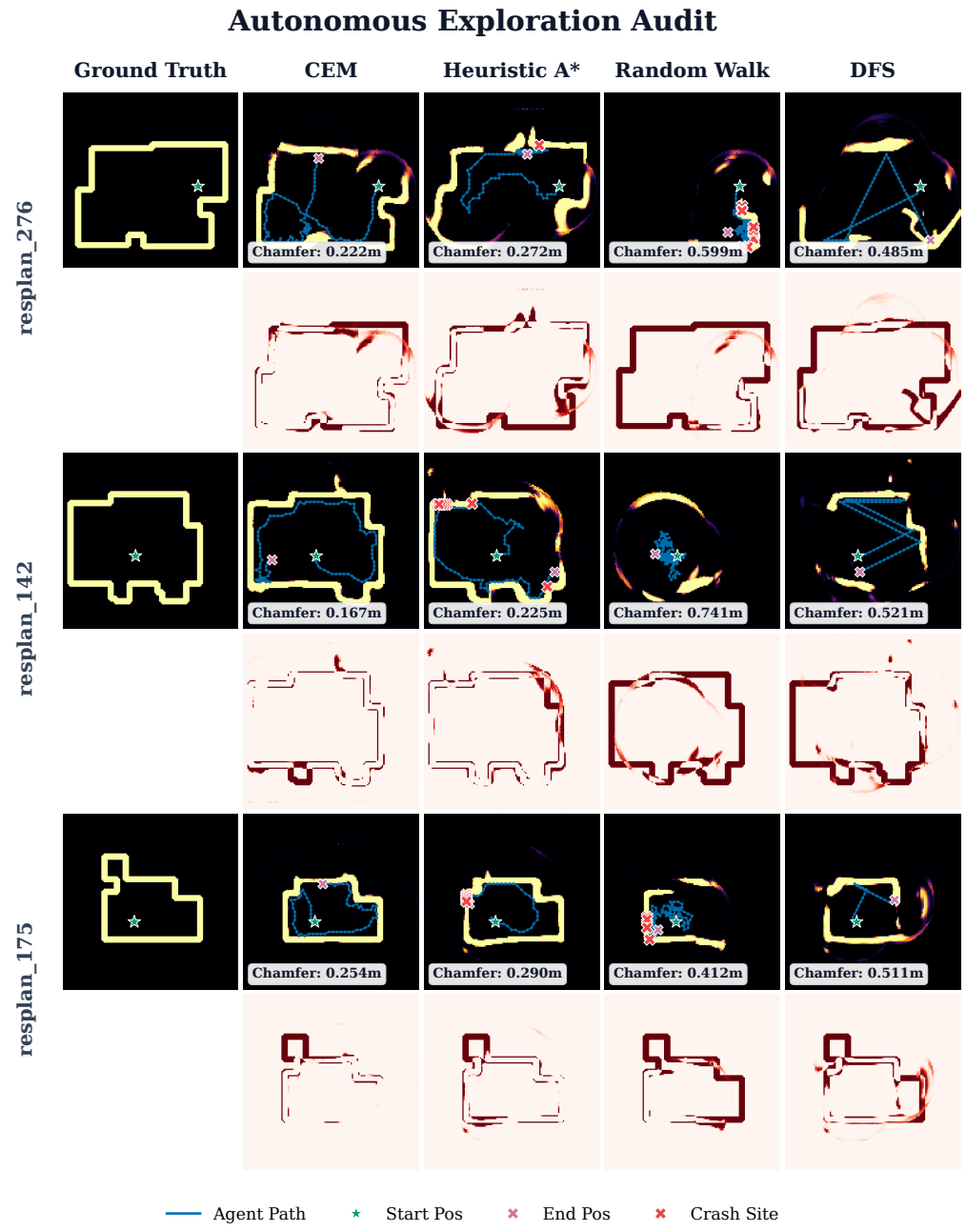


Figure 12: Qualitative closed-loop active acoustic mapping comparison on held-out scenes (continued).

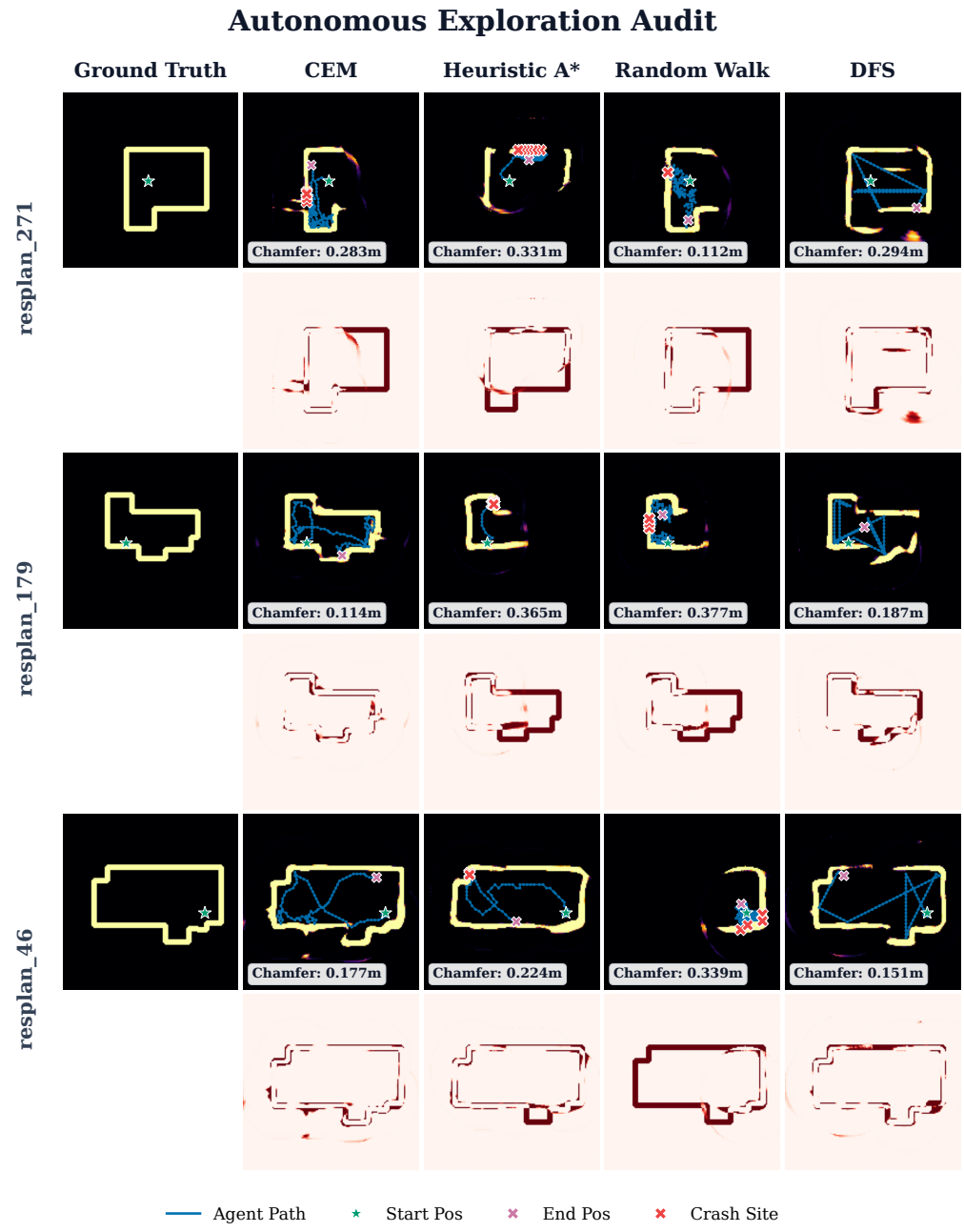


Figure 12: Qualitative closed-loop active acoustic mapping comparison on held-out scenes (continued).

K.2 CEM Rollout Visualizations

We show CEM rollout visualizations to illustrate how the planner evaluates candidate futures.

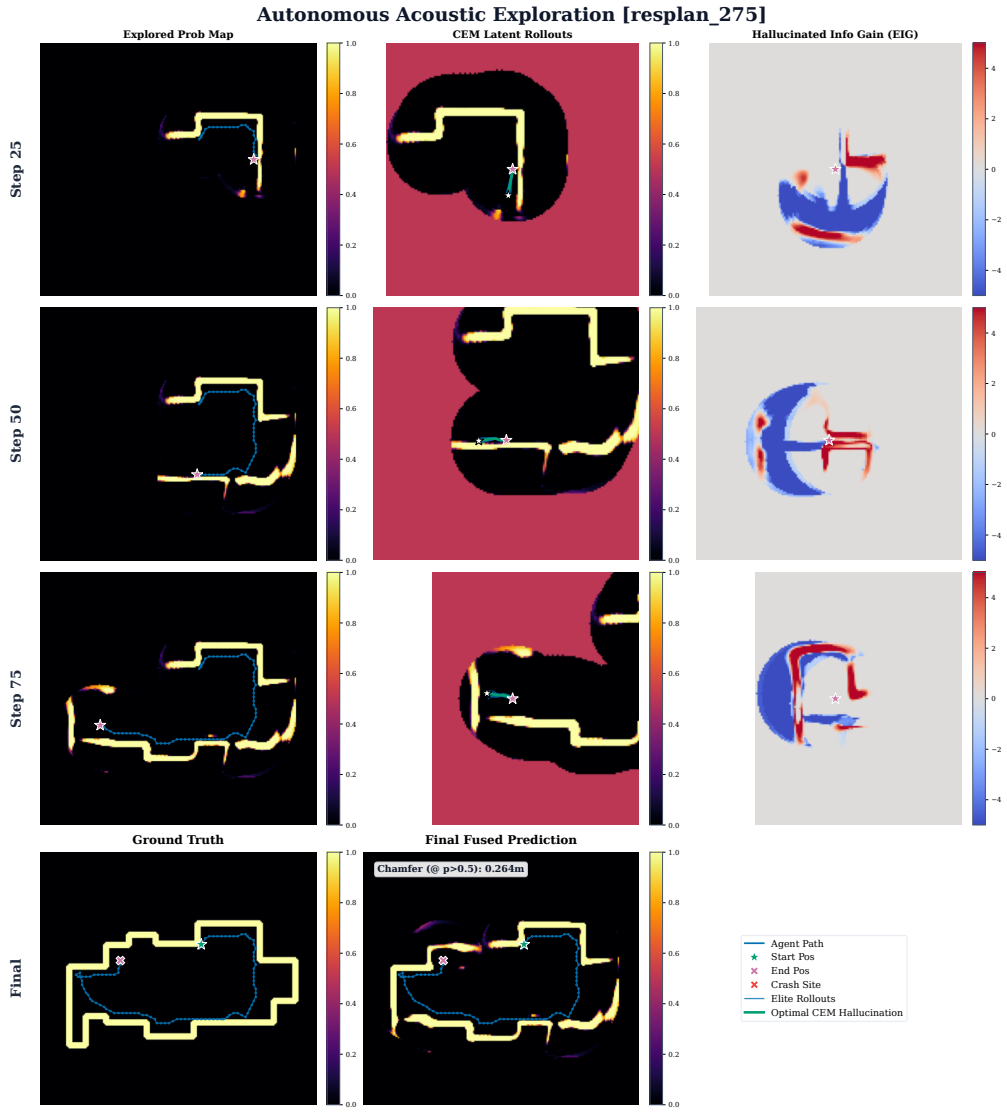


Figure 13: A step-by-step deconstruction of the agent's internal decision-making process during a single zero-shot episode. **Left Column (Explored Prob Map):** The agent's current Bayesian Log-Odds map constructed from the physical acoustic history up to step t. **Center Column (CEM Latent Rollouts):** The internal state of the Cross-Entropy Method planner. The world model predicts dozens of multi-step kinematic futures, evaluates their geometric validity, isolates the elite trajectories (blue), and selects the optimal path (green) entirely within the latent space. **Right Column (Hallucinated Info Gain):** The predicted Bayesian Expected Information Gain (EIG) associated with the optimal rollout. Blue regions indicate spaces the world model expects to solidify into free space. Red indicates expected wall discoveries.

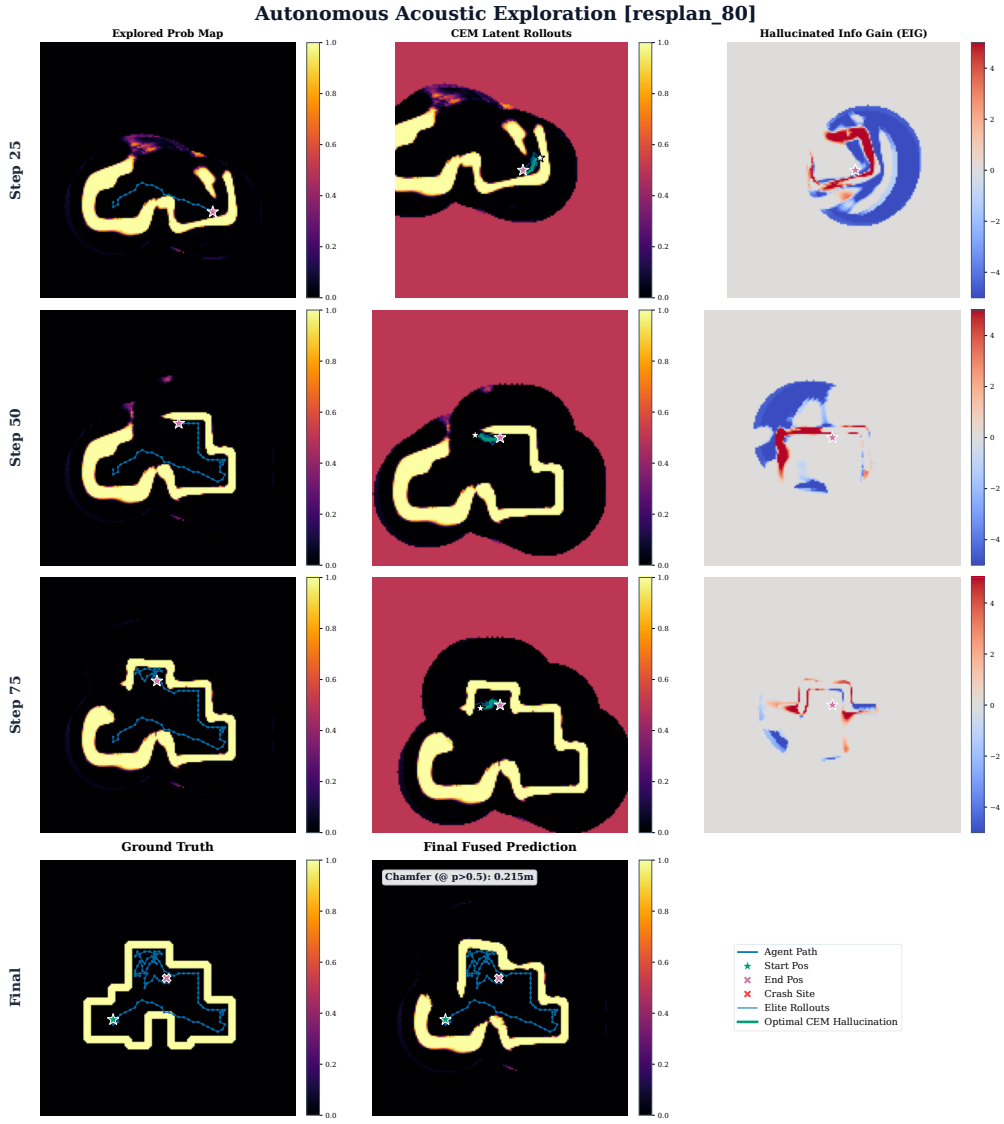


Figure 13: Qualitative CEM rollout visualizations on held-out scenes (continued).

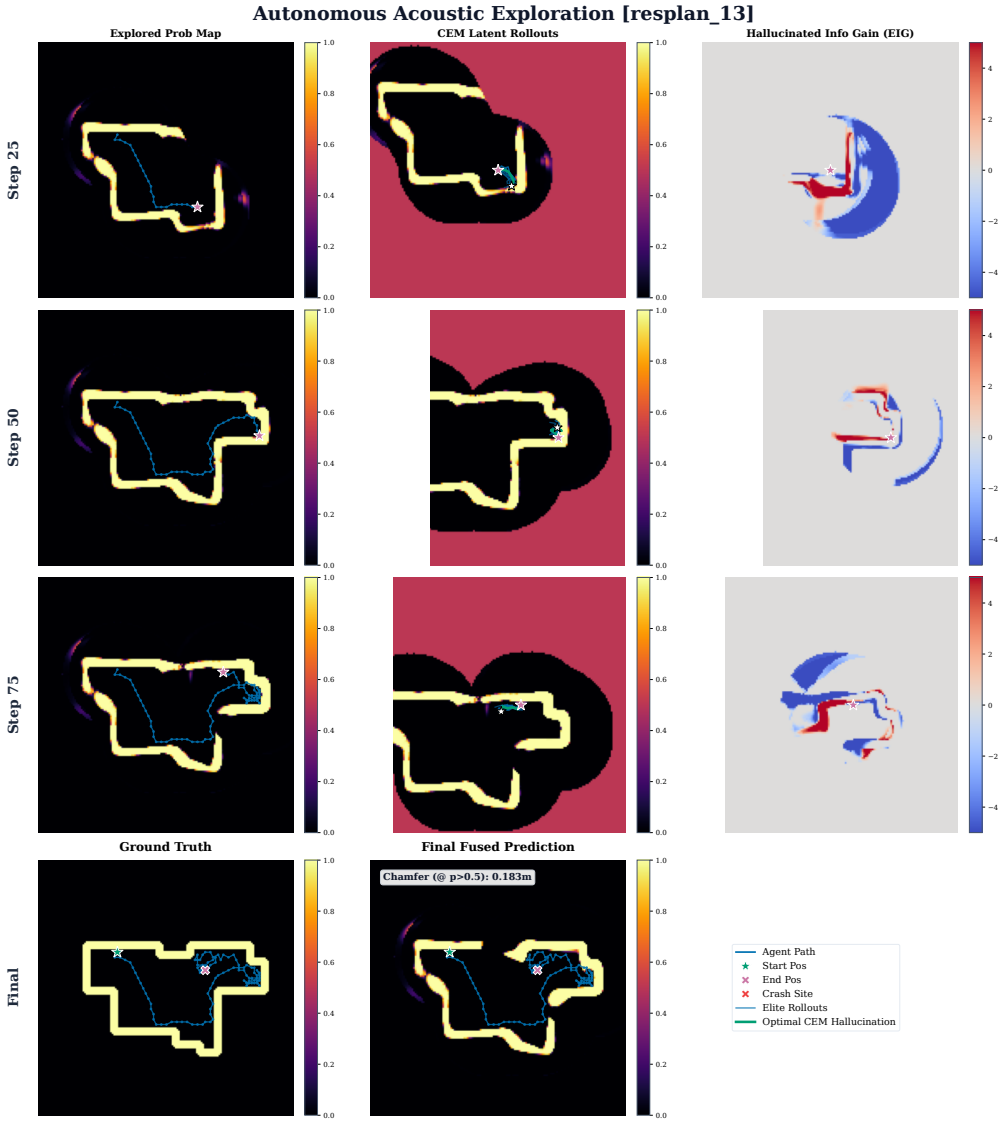


Figure 13: Qualitative CEM rollout visualizations on held-out scenes (continued).

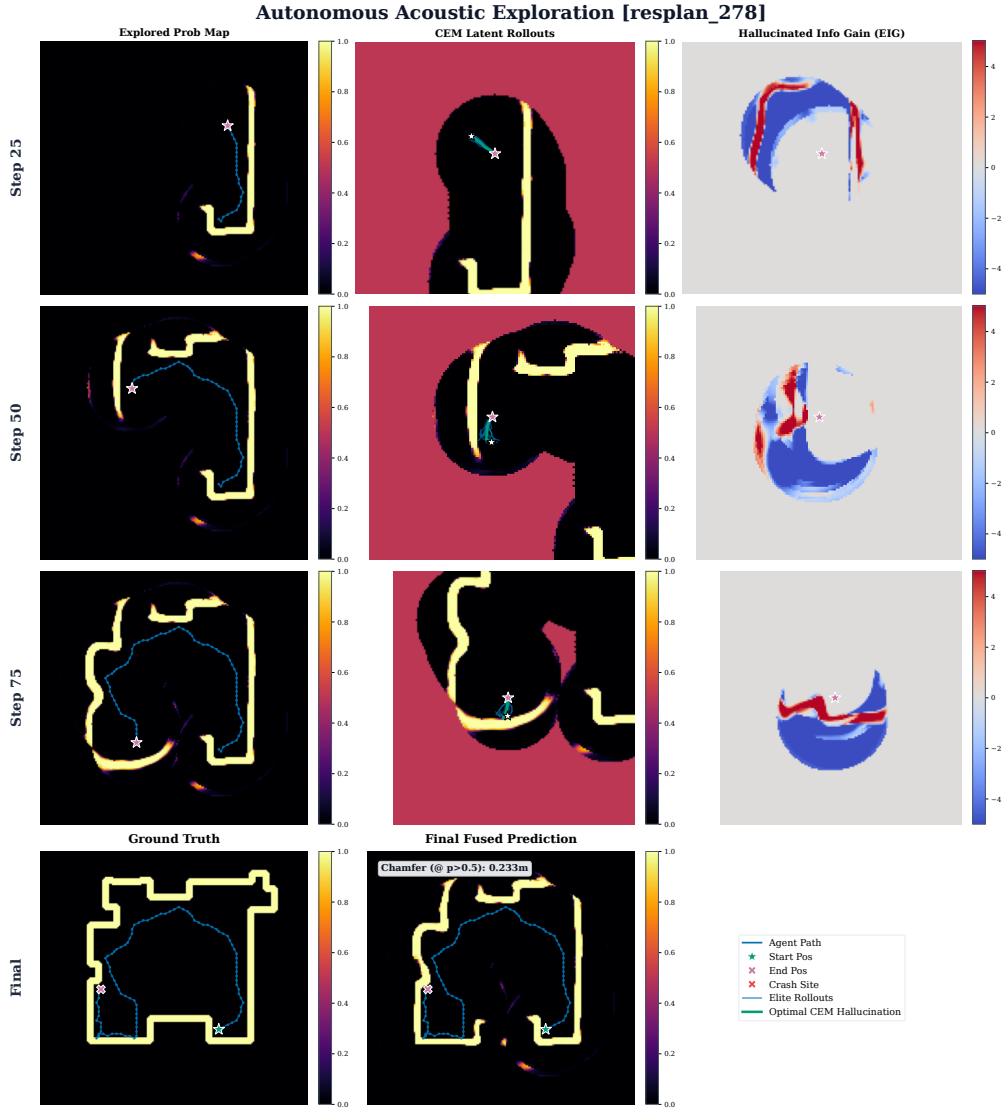


Figure 13: Qualitative CEM rollout visualizations on held-out scenes (continued).

L LATENT SPACE VISUALIZATIONS

This appendix probes the internal representations learned by the acoustic world model beyond reconstruction and active-mapping metrics. All analyses use held-out scenes from MP3D, Atlanta, and ResPlan. We examine four aspects of the latent space: scene-level identity, spatial continuity, decoder attention over trajectory steps, and the effect of kinematic-acoustic fusion.

First, we test whether fused world-state latents form scene-level acoustic fingerprints. Second, we ask whether latent similarity varies smoothly with sensor position inside a scene. Third, we inspect which trajectory observations the spatial decoder attends to when reconstructing occupancy. Fourth, we compare blind audio-only latents to fused motion-conditioned latents to quantify how much pose history modifies the acoustic representation.

L.1 PER-SCENE ACOUSTIC FINGERPRINTS

For each dataset (MP3D, Atlanta, ResPlan), we consider held-out scenes with at least $K = 10$ valid source positions. For every scene, we sample $n_{\text{traj}} = 6$ momentum-walk trajectories using a deterministic seed for each scene-trajectory pair, with step size $0.5\,\text{m}$ and length $K = 10$. Each trajectory is passed through the world model using the same per-dataset relative-pose normalization used during training. We collect two representations at every step: the blind acoustic embeddings u_i from the audio-only encoder (*blind* encoder), and the pose-conditioned world-state tokens $S_t[i]$ produced by the dynamics model (Equation (6)). The dynamics model combines, or fuses, the sequence of blind acoustic embeddings with the corresponding relative poses. Thus, $S_t[i]$ reflects both what was heard and where it was heard along the trajectory.

The per-scene dynamics fingerprint is defined as the mean of all $n_{\text{traj}} \cdot K = 60$ world-state tokens collected from that scene. For comparison, the per-scene blind fingerprint is the mean of 64 blind acoustic embeddings sampled from random source positions in the same scene. We compute these fingerprints for up to 250 held-out scenes per dataset. Atlanta contains fewer held-out scenes than this cap, as indicated in the legends. The resulting D-dimensional per-scene vectors are embedded into 2D using UMAP **?** with `n_neighbors`=15, `min_dist`=0.15, and a fixed random seed. We run UMAP separately for the blind and dynamics fingerprints, so the two panels use the same scenes and projection settings, differing only in whether the representation is audio-only or pose-conditioned.

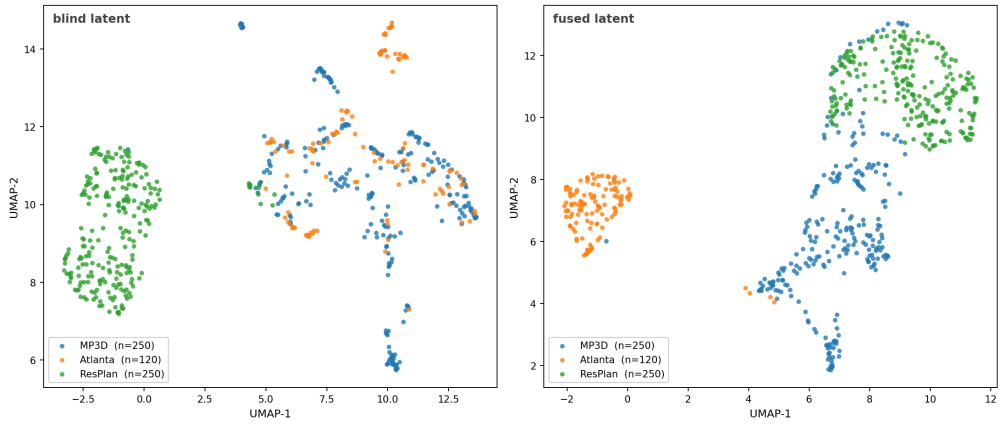
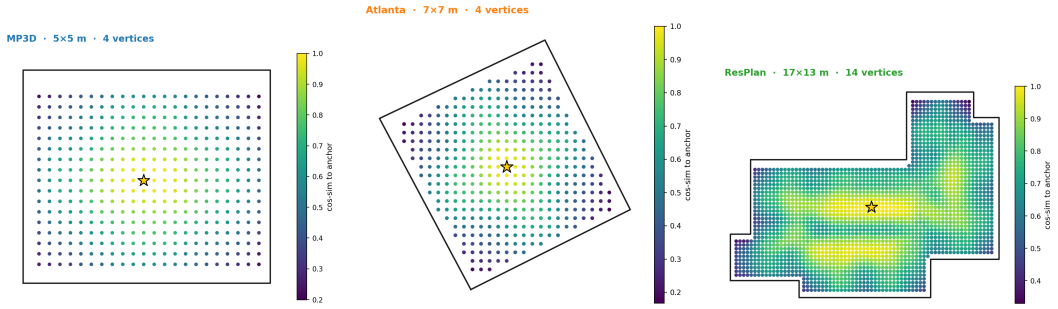


Figure 14: Per-scene acoustic fingerprints. Each marker is one held-out scene, colored by its source dataset. **Left**: UMAP of the mean *blind* latent per scene, computed from the audio-only encoder before pose conditioning. **Right**: UMAP of the mean *dynamics* world-state latent per scene, computed after the dynamics model fuses blind acoustic embeddings with relative pose information along trajectories, using the same scenes and UMAP hyperparameters. The datasets separate more clearly after the dynamics model than in the blind acoustic space, indicating that trajectory-level pose-conditioned acoustic aggregation makes scene identity more discriminative than raw RIR content alone. ResPlan apartment-scale layouts and the single-room MP3D / Atlanta scenes occupy distinct regions in the dynamics panel, consistent with systematic differences in scale and topology. Together, the two panels suggest that the learned world state captures scene-level geometric structure rather than collapsing to a generic acoustic prior.

L.2 Latent Geodesics on the Polygon

To probe the spatial continuity of the audio-only acoustic representation, we select an anchor pose in each scene: the source position closest to the room centroid. We then compute the cosine similarity between the anchor blind embedding and the blind embedding at every other valid sensing pose in the same scene. These embeddings are the per-RIR outputs of the blind acoustic encoder e_θ (*blind* encoder). No relative-pose information and no trajectory context enter the computation, so the resulting field reflects what the audio-only encoder alone can resolve about position. Painting the similarity values onto the room polygon reveals whether the blind embedding already forms a smooth spatial code over the scene, or whether it fragments into unrelated local embeddings. We show one representative held-out scene per dataset.



(a) MP3D scene. (b) Atlanta scene. (c) ResPlan scene.

Figure 15: Latent geodesics painted on the room polygon. For each held-out scene, the star marks the anchor source position closest to the room centroid. The heatmap shows cosine similarity between the anchor blind embedding and the blind embedding at every other valid sensing pose, computed from the per-RIR outputs of the blind acoustic encoder e_θ alone (no fuser, no pose conditioning, no trajectory history). Warmer colors indicate higher similarity. Across all three scenes the audio-only similarity already varies smoothly over space rather than forming fragmented patches, indicating that the blind encoder by itself produces a continuous spatial code - a useful baseline against which to measure how much additional spatial structure is contributed by the kinematic-acoustic fuser (see Appendix L.4). The decay pattern reflects scene scale and topology: compact MP3D and Atlanta rooms show a coherent local gradient, while the larger ResPlan apartment exhibits a broader, more gradual similarity field.

L.3 Decoder Cross-Attention

To understand which trajectory observations the spatial decoder relies on when producing a local occupancy map, we instrument the last cross-attention layer of the spatial query decoder. Each query token (one per output map cell) attends over the K world-state tokens corresponding to the trajectory steps. Averaging the attention weights across all output query tokens and over heads yields one mean attention weight per trajectory step, which we visualize both on the polygon and as a bar chart.

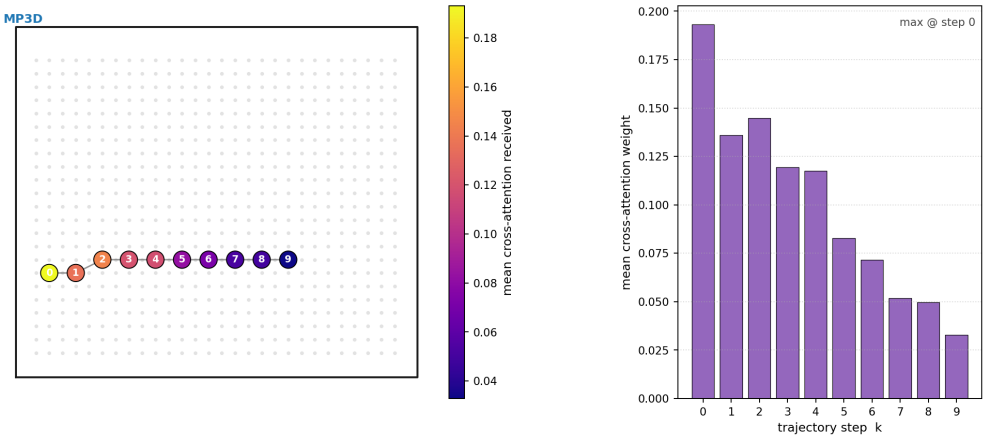


Figure 16: Decoder cross-attention on a representative held-out trajectory ($K = 10$). Left: the trajectory steps painted on the polygon, colored by the average cross-attention weight received from the spatial decoder's last cross-attention layer (averaged over all output query tokens). Right: the same attention weights as a bar chart over trajectory step index. The decoder attends most strongly to the *earliest* trajectory steps, with weights decaying nearly monotonically toward the most recent step. This is consistent with implicit acoustic trilateration: the decoder anchors on observations that provide the longest spatial baseline relative to the current pose, rather than on the most recent observation. The pattern is robust across scenes and datasets.

L.4 Blind vs. Dynamics Modulation

The blind acoustic encoder produces an audio-only latent for each RIR. The dynamics model then fuses this acoustic latent with the trajectory of relative poses, producing a pose-conditioned world-state token. To quantify how strongly the dynamics model transforms the audio-only representation, and how this transformation evolves over a trajectory, we compute the L2 distance between the blind acoustic latent and the corresponding dynamics world-state token at each step of a representative held-out MP3D trajectory:

$$\|S_t[k] - u_k\|_2 .$$

Here, u_k is the blind acoustic embedding for the RIR at step k, and $S_t[k]$ is the corresponding world-state token after the dynamics model has combined acoustic and relative-pose information.

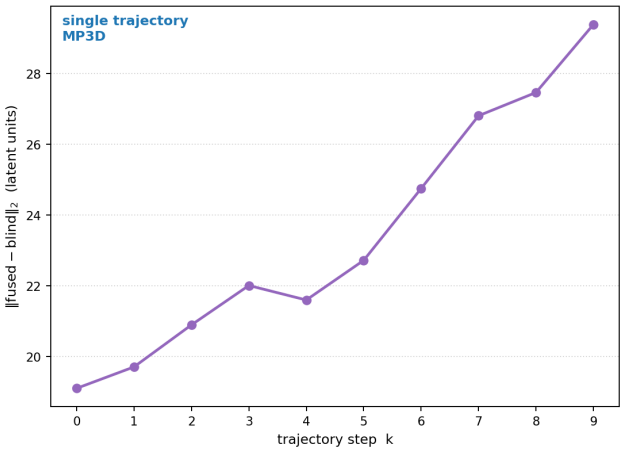


Figure 17: Blind-to-dynamics modulation magnitude along a single held-out MP3D trajectory ($K = 10$). At each trajectory step k, we plot $\|S_t[k] - u_k\|_2$, the L2 distance between the audio-only blind embedding and the corresponding pose-conditioned world-state token. The modulation is large at every step ($\sim$19-29 latent units), showing that the dynamics model substantially transforms the acoustic representation rather than applying a small correction. The modulation also grows nearly monotonically with k, indicating that the world-state tokens become increasingly shaped by accumulated trajectory context. Together with Appendix L.3, this suggests a division of labor between the dynamics model, which builds pose-aware acoustic features over time, and the decoder, which reads out geometry from the resulting world state.

M Planning Budget Ablation

This separate budget sweep compares World-Model CEM with Random Walk, DFS, and Frontier + A*. It illustrates the reconstruction and collision trade-off across sensing budgets. The matched CEM and oracle comparisons are reported in Table 2.

Table 15: Closed-loop active acoustic mapping on **ResPlan** held-out test scenes under a fixed budget of 25 environment steps. Threshold-dependent reconstruction metrics use occupancy threshold 0.5. Values are mean ± standard deviation across test episodes. CD and DF denote Chamfer and DF-MAE (m), SER is in m^2/m, and crash rate is per 100 m. Bold and underlining mark the best and second-best non-oracle means.

Planner	Exploration / Stability			Geometric Error		ϵ-F1		
	Coverage (%) ↑	SER (m^2/m) ↑	Crash/100m ↓	Chamfer (m) ↓	DF-MAE (m) ↓	$F1_{\epsilon=0.15\mathrm{m}}$ ↑	$F1_{\epsilon=0.30\mathrm{m}}$ ↑	$F1_{\epsilon=0.45\mathrm{m}}$ ↑
Random Walk	42.27 ± 13.57	9.97 ± 0.83	5.15 ± 11.07	0.402 ± 0.217	0.402 ± 0.199	0.304 ± 0.172	0.662 ± 0.162	0.774 ± 0.128
DFS Heuristic	55.52 ± 17.44	$\mathbf{17.42 \pm 13.37}$	$\mathbf{0.12 \pm 1.23}$	0.427 ± 0.197	0.404 ± 0.171	0.297 ± 0.147	0.657 ± 0.142	0.764 ± 0.115
Frontier + A*	$\mathbf{65.63 \pm 18.79}$	15.89 ± 10.85	4.20 ± 17.46	0.328 ± 0.162	0.333 ± 0.168	0.314 ± 0.136	0.702 ± 0.137	0.817 ± 0.103
World-Model CEM (Ours)	64.81 ± 17.70	15.38 ± 6.85	0.30 ± 2.08	$\mathbf{0.302 \pm 0.148}$	$\mathbf{0.291 \pm 0.137}$	$\mathbf{0.343 \pm 0.147}$	$\mathbf{0.746 \pm 0.127}$	$\mathbf{0.839 \pm 0.095}$

Table 16: Closed-loop active acoustic mapping on **ResPlan** held-out test scenes under a fixed budget of 50 environment steps. Threshold-dependent reconstruction metrics use occupancy threshold 0.5. Values are mean ± standard deviation across test episodes. CD and DF denote Chamfer and DF-MAE (m), SER is in m^2/m, and crash rate is per 100 m. Bold and underlining mark the best and second-best non-oracle means.

Planner	Exploration / Stability			Geometric Error		ϵ-F1		
	Coverage (%) ↑	SER (m^2/m) ↑	Crash/100m ↓	Chamfer (m) ↓	DF-MAE (m) ↓	$F1_{\epsilon=0.15\mathrm{m}}$ ↑	$F1_{\epsilon=0.30\mathrm{m}}$ ↑	$F1_{\epsilon=0.45\mathrm{m}}$ ↑
Random Walk	48.51 ± 14.87	5.80 ± 0.56	6.18 ± 9.75	0.392 ± 0.214	0.383 ± 0.190	0.286 ± 0.139	0.669 ± 0.141	0.783 ± 0.117
DFS Heuristic	65.90 ± 20.84	$\mathbf{13.86 \pm 14.42}$	$\mathbf{0.21 \pm 1.05}$	0.395 ± 0.175	0.370 ± 0.149	0.321 ± 0.145	0.678 ± 0.125	0.787 ± 0.094
Frontier + A*	$\mathbf{84.22 \pm 17.49}$	12.69 ± 11.53	10.91 ± 38.71	0.245 ± 0.127	0.242 ± 0.120	0.363 ± 0.133	0.777 ± 0.126	0.879 ± 0.092
World-Model CEM (Ours)	83.27 ± 16.16	10.95 ± 7.33	0.52 ± 2.83	$\mathbf{0.237 \pm 0.127}$	$\mathbf{0.227 \pm 0.110}$	$\mathbf{0.409 \pm 0.144}$	$\mathbf{0.813 \pm 0.105}$	$\mathbf{0.886 \pm 0.077}$

Table 17: Closed-loop active acoustic mapping on **ResPlan** held-out test scenes under a fixed budget of 75 environment steps. Threshold-dependent reconstruction metrics use occupancy threshold 0.5. Values are mean ± standard deviation across test episodes. CD and DF denote Chamfer and DF-MAE (m), SER is in m^2/m, and crash rate is per 100 m. Bold and underlining mark the best and second-best non-oracle means.

Planner	Exploration / Stability			Geometric Error		ϵ-F1		
	Coverage (%) ↑	SER (m^2/m) ↑	Crash/100m ↓	Chamfer (m) ↓	DF-MAE (m) ↓	$F1_{\epsilon=0.15\mathrm{m}}$ ↑	$F1_{\epsilon=0.30\mathrm{m}}$ ↑	$F1_{\epsilon=0.45\mathrm{m}}$ ↑
Random Walk	54.49 ± 16.74	4.30 ± 0.44	4.96 ± 6.73	0.375 ± 0.186	0.359 ± 0.174	0.300 ± 0.138	0.684 ± 0.132	0.793 ± 0.104
DFS Heuristic	74.24 ± 20.14	11.61 ± 14.51	**0.15 ± 0.67**	0.361 ± 0.159	0.341 ± 0.128	0.329 ± 0.130	0.709 ± 0.120	0.807 ± 0.097
Frontier + A*	**90.89 ± 15.04**	**11.79 ± 11.77**	16.83 ± 63.27	0.212 ± 0.125	0.214 ± 0.116	0.383 ± 0.131	0.807 ± 0.125	**0.905 ± 0.089**
World-Model CEM (Ours)	90.12 ± 14.24	8.77 ± 7.58	0.51 ± 1.69	**0.210 ± 0.113**	**0.206 ± 0.096**	**0.425 ± 0.148**	**0.840 ± 0.100**	0.902 ± 0.075

Table 18: Closed-loop active acoustic mapping on **ResPlan** held-out test scenes under a fixed budget of 150 environment steps. Threshold-dependent reconstruction metrics use occupancy threshold 0.5. Values are mean ± standard deviation across test episodes. CD and DF denote Chamfer and DF-MAE (m), SER is in m^2/m, and crash rate is per 100 m. Bold and underlining mark the best and second-best non-oracle means.

Planner	Exploration / Stability			Geometric Error		ϵ-F1		
	Coverage (%) ↑	SER (m^2/m) ↑	Crash/100m ↓	Chamfer (m) ↓	DF-MAE (m) ↓	$F1_{\epsilon=0.15\mathrm{m}}$ ↑	$F1_{\epsilon=0.30\mathrm{m}}$ ↑	$F1_{\epsilon=0.45\mathrm{m}}$ ↑
Random Walk	64.62 ± 18.57	2.64 ± 0.30	5.31 ± 5.85	0.354 ± 0.174	0.319 ± 0.137	0.275 ± 0.123	0.676 ± 0.123	0.800 ± 0.098
DFS Heuristic	78.59 ± 22.61	10.89 ± 15.36	**0.37 ± 2.36**	0.329 ± 0.170	0.317 ± 0.138	0.347 ± 0.139	0.737 ± 0.126	0.830 ± 0.102
Frontier + A*	93.99 ± 13.31	**11.17 ± 12.01**	33.14 ± 146.90	**0.193 ± 0.119**	**0.194 ± 0.100**	0.394 ± 0.128	0.826 ± 0.123	**0.921 ± 0.087**
World-Model CEM (Ours)	**96.36 ± 9.72**	5.78 ± 7.94	1.19 ± 8.52	0.198 ± 0.110	0.196 ± 0.085	**0.428 ± 0.147**	**0.845 ± 0.105**	0.911 ± 0.073

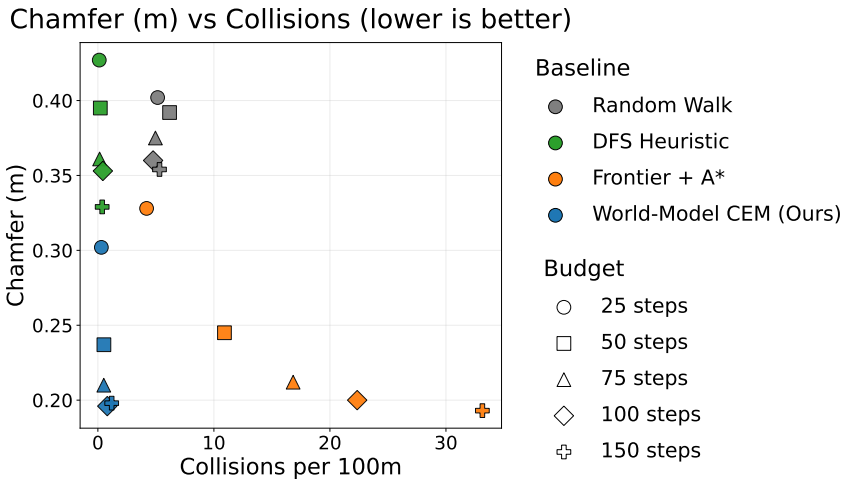


Figure 18: Reconstruction accuracy versus collision rate across sensing budgets. Each point corresponds to one planner at a fixed episode budget, with marker shape indicating the number of steps. Lower-left is better. World-Model CEM consistently occupies the low-collision, low-error regime, while Frontier + A* shows increasing collision rates as the budget grows.

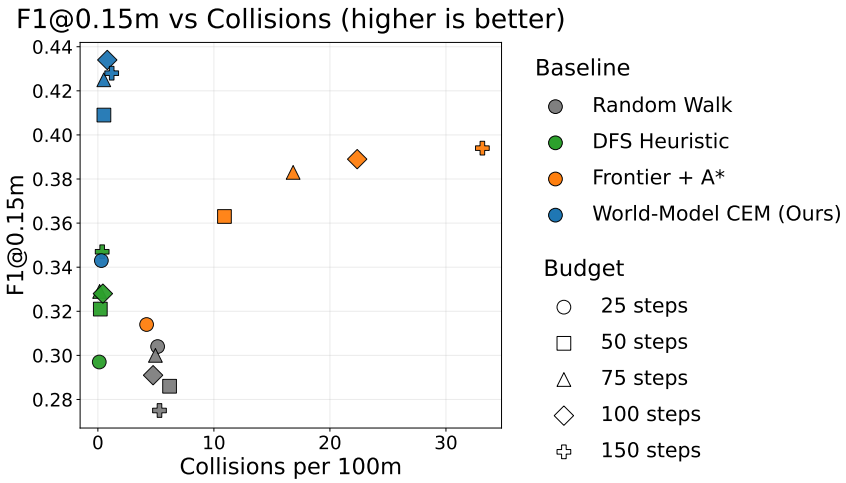


Figure 19: Boundary F1 at 0.15 m versus collision rate across sensing budgets. Higher F1 and lower collision rate are preferable. Marker shape indicates the episode budget.

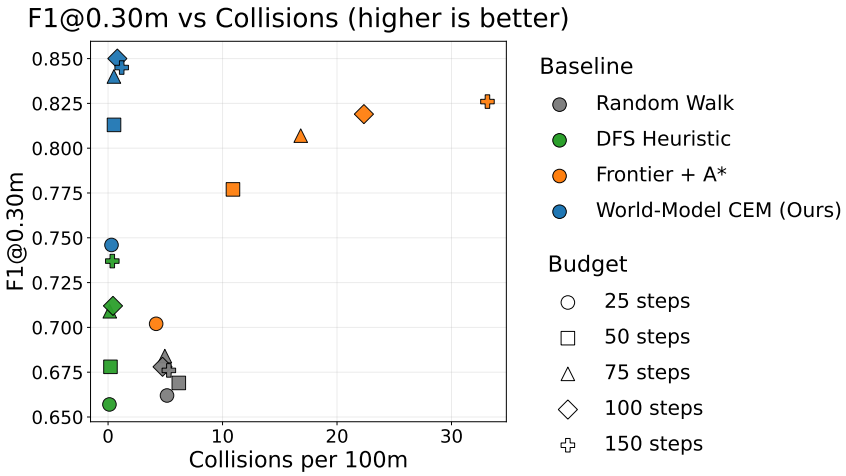


Figure 20: Reconstruction accuracy versus collision rate across sensing budgets.

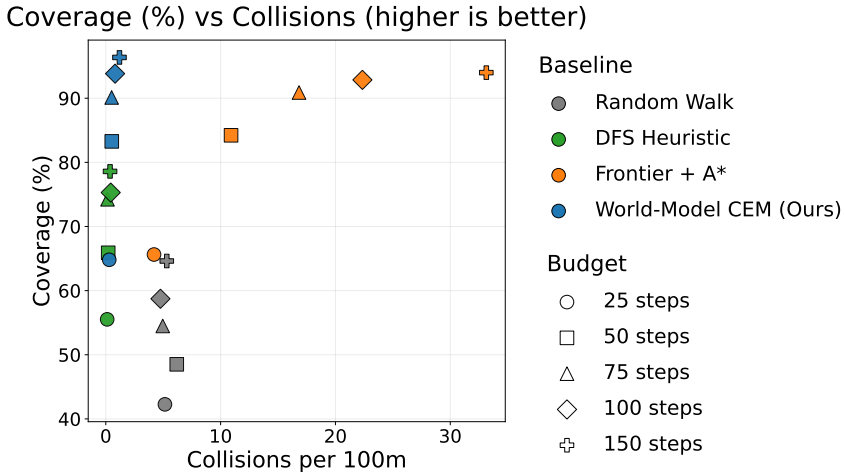


Figure 21: Coverage versus collision rate across sensing budgets.

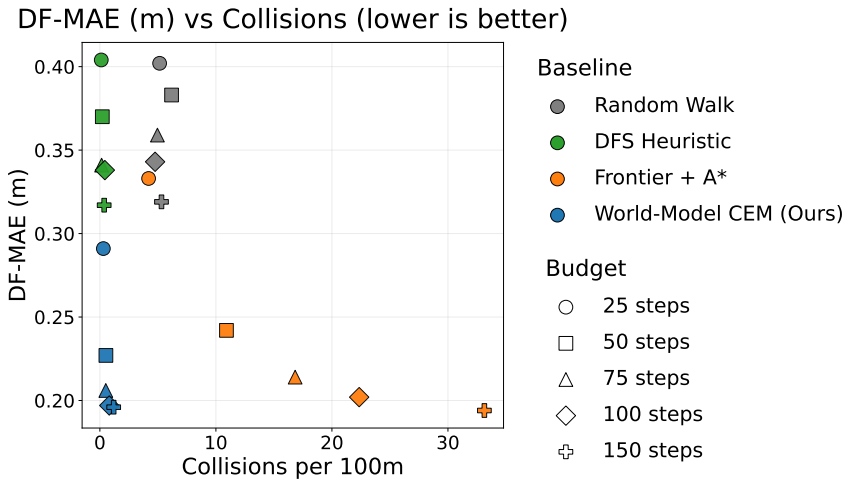


Figure 22: Reconstruction accuracy versus collision rate across sensing budgets for DF-MAE.