Loom: Leveraging In-situ Smart Speakers for Scalable Neural Floorplan Inference

Anonymous Author(s)

ABSTRACT

Accurate indoor floorplans are foundational for emerging smart home applications. Yet, acquiring this geometry typically relies on intrusive dedicated hardware or active crowdsourced mobile scanning, rendering widespread adoption impractical. In this paper, we present Loom, the first neural floorplan inference system that recovers room geometry using in-situ, commodity smart speakers without any active user intervention. However, translating sparse, stationary acoustic signals into geometric boundaries is a highly ambiguous, ill-posed inverse problem. Loom breaks this physical barrier through three core innovations. First, we formulate the layout reconstruction as a physics-guided conditional generation task. At its core, we design a proxy network to model acoustic propagation and constrain the structural search space. Second, we opportunistically reuse ambient echoes from daily user-device interactions as dynamic sound sources, unlocking multi-view spatial parallax without extra burden. Third, we employ a self-evolving mechanism to seamlessly adapt to unlabeled, heterogeneous room semantics out-of-the-box. Extensive evaluations show that Loom achieves an SSIM of 0.83 in furnished rooms. We believe Loom will pave the way for the ubiquitous spatial intelligence.

1 INTRODUCTION

As our homes undergo a remarkable transformation towards *spatial intelligence* [29, 37], future smart systems are expected to move beyond simple command execution to *perceive*, *interpret*, and perform *spatial reasoning* within their physical context [7, 56, 65]. For instance, when a user simply commands "turn on the light," a truly intelligent assistant should localize the direction of the voice and cross-reference it with the physical room layout to determine the zone, eliminating the need for rigid device labels. Central to realizing this vision is an awareness of the indoor geometry, specifically the floorplan, which provides the essential map and coordinate system for all context-aware interactions.

Despite this necessity, one striking paradox is that the current computing infrastructure in smart homes is spatially blind. Fundamentally, this blindness stems from the fact that these IoT devices are not equipped with the ability to sense the space, remaining oblivious to the geometric environment they inhabit. This leads to an isolated ecosystem where floorplan acquisition is completely decoupled from daily perception performed by in-situ devices. Concretely, current approaches for acquiring room geometry rely predominantly on two routes: dedicated instrumentation such as LiDARs [8, 46, 49, 50, 87] and depth cameras [34, 43], or crowdsourced mobile scanning [26, 58, 101, 102]. However, both routes present major obstacles to widespread adoption. On one hand, the integration of specialized mapping hardware incurs prohibitive costs, and deploying such dedicated sensors in private domestic spaces is considered *intrusive*. On the other hand, crowdsourced approaches are difficult to scale, as they often rely on occupants' *active* cooperation to manually perform dense sweeps [20, 84, 91]. Such reliance on user-initiated, time-consuming actions inevitably leads to massive friction in user compliance. Consequently, there is an urgent need for a scalable solution that can recover floorplans automatically, without requiring additional hardware or excessive user intervention.

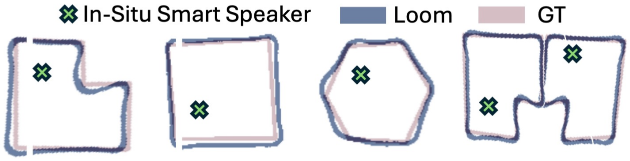


Figure 1: Loom repurposes the in-situ smart speaker for floormap inference.

To build such a solution, we must reuse an existing infrastructure that has already woven into our domestic fabric. Fortunately, we observe an opportunity in smart speakers (*e.g.*, Amazon Echo, Google Home), which have emerged as the *de facto* central control hubs [28]. It is observed that over 57% of U.S. households own at least one smart speaker [72], where nearly half will own more than one device [18]. In other words, smart speakers have become virtually ubiquitous and accessible indoor infrastructure, offering a unique platform to infer room geometry without introducing new hardware. Meanwhile, among various modalities like Wi-Fi, acoustic strikes a great balance between resolution, ubiquity, and multipath reflections, rendering it a promising modality for capturing the spatial semantics indoors.

Motivated by these, we naturally ask: *Can we decode the floor map solely via the in-situ smart speakers?* In this paper, we propose Loom, the first scalable neural floorplan inference system leveraging pervasive in-situ smart speakers. Unlike active crowdsourcing, Loom opportunistically reuses the sound recordings from COTS smart speakers with minimal user intervention. However, transforming a set of passive acoustic signals into an accurate floorplan is a highly underdetermined and indirect problem, which embodies multiple challenges as follows:

■ **Ambiguity in Structural Inference:** Transforming acoustic observations into an accurate floorplan inevitably embodies an ill-posed inverse problem. Unlike vision sensors that capture explicit spatial semantics, a microphone collapses intricate spatial interactions into a highly compressed 1D temporal recording. While advanced neural methods [14, 45, 51, 52, 100] can enable us to instead view the full acoustic responses as a high-dimensional encoding of the whole space, these methods primarily focus on the forward mapping, *i.e.*, learning to synthesize a signal from a known space, leaving the geometric information implicitly entangled. In fact, the mapping from acoustics to geometry is intrinsically *one-to-many*: a single acoustic echo could be explained by many possible layouts. This inversion introduces a deep ambiguity and intrinsic uncertainty absent in forward field synthesis. Without explicit structural priors to bound a valid solution space, naive optimization or search over a boundless continuous space is computationally infeasible. To address this, we formulate the reconstruction task as a physics-guided conditional generation task. Specifically, we decompose the complicated inverse problem into modeling of two complementary distributions, *i.e.*, acoustic physics likelihood and geometry prior. To learn spatial semantics, we design a proxy network that maps acoustic signals to explicit proxy conditions. Concurrently, we utilize a latent diffusion network to learn the structural priors of valid indoor geometries. By constraining the solution space to a low-dimensional manifold, Loom effectively eliminates structurally invalid layouts and confines the search space.

■ **Scarcity of Diverse Spatial Parallax:** Traditional reconstruction methods predominantly rely on actively moving sensors, such as handheld scanning [58], spinning radars [42], or mobile robots [49, 88], to traverse a space and thus provide the essential parallax required to triangulate boundaries and resolve occlusions. Conversely, in-situ smart speakers are usually tied to a fixed location. This physical immobility means that the device inherently lacks a global perspective and thus fails to provide sufficient evidence to reliably deduce the global arrangement of complex rooms. The challenge, therefore, lies in how to overcome this hardware immobility and acquire necessary spatial perspectives without requiring users to physically relocate the device. Our key insight arises from the natural functionality of smart speakers, *i.e.*, interacting with users. Driven by this, we opportunistically harvest the ambient human-speaker interaction. We utilize continuous human interactions in daily routines as mobile and dynamic sound sources, transforming spatio-temporal human-speaker interactions into rich spatial parallax. By doing so, we obtain the necessary multi-view acoustic perspectives to pinpoint the space boundaries.

■ **Adaptability to Heterogeneous Environments:** Real-world residential environments are vastly heterogeneous. Different room functions and shapes dictate radically different acoustic propagation patterns. A monolithic neural network often fails to generalize across such diverse spatial semantics. More critically, in practice, we have no access to ground-truth floorplans for supervised fine-tuning or post-hoc calibration. Therefore, Loom faces a dual-fold scalability challenge. First, it should decouple and adapt to the specific acoustic characteristics of varying room types out-of-the-box. Second, without explicit user annotations, the model must continuously refine its geometric parameters driven solely by the physical consistency of ambient acoustic interactions. To achieve zero-calibration adaptation, we design a systematic framework. First, we employ a Mixture-of-Experts (MoE) architecture that routes acoustic features to the corresp based on semantic room conditions, benefiting from the expert knowledge of different geometry priors. Second, we exploit the inherent acoustic tracking capabilities of modern smart speakers. By extracting intermediate acoustic properties, including the Angle of Arrival (AoA) of the user and Room Impulse Responses (RIRs), as pseudo-labels, Loom enables a self-evolving mechanism that continually fine-tunes the geometric parameters, which allows the system to adapt to entirely unseen, unlabeled domestic spaces seamlessly.

We conduct extensive experiments to demonstrate the effectiveness of Loom. In fully furnished rooms, Loom achieves an average SSIM of 0.83 and an IoU of 0.65, outperforming existing baselines by 169.3% and 65.0%, respectively. Furthermore, it yields an average Chamfer distance of merely 1.25m, lowering the geometric reconstruction error by 87.13%. Moreover, Loom exhibits strong generalization, maintaining an SSIM of 0.77 and an IoU of 0.48 in entirely unseen environments. We validate that Loom is highly consistent across diverse practical scenarios. Meanwhile, our online adaptation boosts the performance by 88.5% in real-world scenarios.

Contributions: We summarize our contributions below.

❶ Loom is the first system that leverages the in-situ smart speakers for scalable neural floormap inference with minimal human intervention.

❷ We translate the problem into a physics-guided conditional generation problem and decompose the ill-posed inverse problem into learnable components. We design a multi-task proxy network to model the acoustic propagation and a physics-guided latent diffusion model to generate the floormap.

❸ We implement Loom as a practical system and conduct extensive evaluation to show its efficacy. As a side product, we develop an acoustic-based ray tracing engine, named ARTrace, that completely complies with the acoustic physics to support the data collection.

2 PROBLEM SCOPE

Applications: Floorplan serves as a geometric foundation for higher-order spatial intelligence through different capabilities, as illustrated in Fig. 2. First, it enables *spatial diagnostics*, allowing Wi-Fi systems to identify physical root causes rather than offering generic, context-agnostic advice. Second, it facilitates finer-grained *control*, where HVAC systems can adapt airflow based on room geometry to address specific localized needs. Finally, it fosters ecosystem *collaboration* by resolving ambiguity of the sensing systems. For example, the system can infer the direction of a user's gesture and seamlessly coordinate operations across multiple appliances.
From Crowdsourcing to Ambient Sensing: While capturing spatial geometry is vital, existing solutions [21, 26, 58] predominantly dictate an *active crowdsourcing* paradigm. They require users to deliberately walk through their homes while waving their devices to collect dense spatial measurements. We argue that this approach suffers from scalability issues for two key reasons. First, users are required to perform a dense scanning of the whole room, which is time-consuming and intrusive. More importantly, such a formulation tailors user-centric action, which is usually infeasible in real-world applications.

To address this, Loom reframes the floormap reconstruction task into a new problem space. Instead of relying on excessive user scanning, we explore a new paradigm that leverages the in-situ smart speakers to infer the spatial geometry. This brings three key benefits: ❶ **Zero Hardware Cost:** We repurpose existing pervasive infrastructure into an inference engine of the floormap, without requiring additional new hardware. ❷ **Zero-Effort Operation:** Loom operates passively in the background. Without requiring users to perform tedious sweeping motions or traverse the space, we shift the burden of spatial perception from users or robotics to in-situ smart speakers. ❸ **Non-Visual Modality:** Compared with vision-based SLAM, we employ a privacy-preserving alternative that has been widely adopted and woven into our daily lives for obtaining the geometry.

At a high level, Loom can be deployed as an over-the-air update on existing smart speakers without any hardware modifications. It achieves passive convergence by continuously iterating through daily user interactions and ambient acoustic events of interest. We envision a "zero-effort" deployment model that Loom is exposed as a middleware API that empowers spatial intelligence, providing a foundational interface that bootstraps an embodied AI ecosystem.
Limitations of Existing Work: Notably, Loom faces a different problem space compared with existing approaches. To appreciate it, we first clarify why these works cannot be trivially adopted. ❶ *Neural-Field Rendering:* Recent advances in neural implicit representations have shown promise in modeling wireless/sound fields [45, 51, 52, 77, 100]. However, these methods fundamentally rely on a known global coordinate system to perform grid sampling and positional encoding, which is not even available in our problem settings. Instead, we only possess the relative acoustic measurements. Moreover, these methods struggle to generalize across unseen environments, as the underlying mechanism is to map the coordinates to the signal field, rendering it incapable of discovering the geometry of any unknown environments. ❷ *Differentiable Ray Tracing (DRT):* While DRT has been widely used to optimize acoustic parameters [11, 32, 36, 64], it strictly requires a priori room boundaries to initialize the rays, compute interactions and backpropagate gradients. However, in our problem scope, the boundaries are the unknown information we aim to reconstruct. ❸ Alternatively, many existing works benefit from rich priors, including 3D point clouds [46, 66], and RGB images [12]. Conversely, Loom does not incorporate any additional priors and operates only on acoustic modalities. Consequently, the system must learn to untangle complex spatial semantics merely from highly multiplexed, sparse acoustic echoes. To this end, we have to design a new solution to resolve the problem.

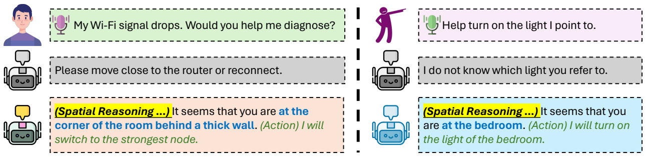
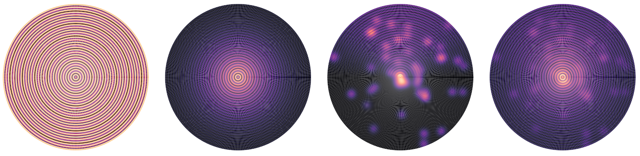

My Wi-Fi signal drops. Would you help me diagnose?
Please move close to the router or reconnect.
(Spatial Reasoning ...) It seems that you are at the corner of the room behind a thick wall. (Action) I will switch to the strongest node.
Help turn on the light I point to.
I do not know which light you refer to.
(Spatial Reasoning ...) It seems that you are at the bedroom. (Action) I will turn on the light of the bedroom.

Figure 2: Illustrations of Spatial Reasoning



(a) Reflection (b) Absorption (c) Diffusion (d) Combined

Figure 3: Visualization of the Acoustic Wave Field.

3 BACKGROUND

The core problem lies in the bidirectional mapping between acoustic propagation and spatial information. In this section, we need to review how sound propagates in the room and how we practically represent the propagation.
Reflection, Diffusion, Absorption: Indoor acoustic propagation primarily includes reflection, diffusion, and absorption [41, 53], as in Fig. 3. When an acoustic wave encounters a surface, a portion of its energy is absorbed, while the remaining energy is reflected back into the space. On smooth boundaries, the wave undergoes deterministic specular reflection. Meanwhile, a significant portion of the acoustic energy undergoes scattering into various angles. The structural layout is thus intrinsically encoded in the acoustic propagation,

making the rigorous modeling of these dynamics a prerequisite for decoding the spatial floorplan.

Room Impulse Response: Notably, the spatial feature is encapsulated by Room Impulse Response (RIR), *i.e.*,

$$h(t,\theta)=\sum_i\left[\prod_j(1-\alpha_{i,j})\right]\cdot\sum_k S_{i,k}\delta(t-\tau_{ik})\delta(\Theta-\Theta_{ik}). \quad (1)$$

Here i means the path from the speaker to the microphone and j denotes the faces it hits on paths i. $\alpha_{i,j}$ represents the absorption coefficient on the j-th surface along the i-th path. $S_{i,k}$ denotes the scattering coefficient associated with the k-th subpath, while $\tau_{i,k}$ is the arriving delay. We explicitly introduce the Angle of Arrival (AoA) as a key attribute of each propagation path, *a.k.a.*, $\Theta=(\phi,\theta)$. Here ϕ is the azimuth while θ is the elevation.

RIR and Room Geometry: While Eq. 1 implies that an RIR encodes spatial cues, it remains unclear how strongly it reflects room geometry. We therefore conduct a feasibility study that measures RIR sensitivity to structural perturbations at three scales, namely changing the room boundary layout, removing a large obstacle, and removing a small daily object, as shown in Fig. 4(b). We quantify the difference using an acoustic Similarity Matrix between RIRs. As shown in Fig. 4(a), altering the room shape strongly changes dominant multipath propagation, resulting in low similarity of 0.15. Removing a large object affects secondary reflections and occlusions, giving a moderate similarity of 0.55. Removing a small object causes a weaker but still detectable change, retaining a high similarity of 0.92. These results suggest that RIRs entangle multi-scale geometric information and are promising for geometry decoding. However, the RIR still suffers from ambiguity and sparsity problems. In the following section, we will rigorously formulate this problem and present a novel framework to resolve them.

4 LOOM DESIGN

4.1 Problem Formulation

Given the geometry $\Omega=\{\Omega_i\}_{i=1}^{N}$, we define a forward physical operator $\mathcal{W}:\mathbb{R}^N\rightarrow\mathbb{R}^M$, which defines

$$S_{\text{obs}}=\mathcal{W}(\Omega)+\eta, \quad (2)$$

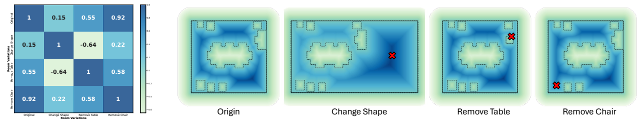
where η is the noise and S_{obs} is the observed signals.

Ambiguity Problem: We start from the ambiguity of the inverse transform of $\mathcal{W}$.

Lemma 1. *From a single point, the inverse operator $\mathcal{W}^{-1}$ is ill-posed. There exists an ambiguity set $\mathcal{E}_{amb}$ that satisfies*

$$\mathcal{W}(\Omega_i)\approx\mathcal{W}(\Omega_j)=S_{obs} \qquad \forall\Omega_i,\Omega_j\in\mathcal{E}_{amb}$$

Proof. In a single-device setup, the degrees of freedom in the scene vastly exceed the independent constraints provided by the sparse signals ($N\gg M$). By linearizing $\mathcal{W}$ at a local geometry Ω_0, we obtain $\mathcal{W}(\Omega_0+\Delta\Omega)\approx\mathcal{W}(\Omega_0)+\mathbf{J}\Delta\Omega$, where $\mathbf{J}$ is the Jacobian matrix of $\mathcal{W}$. Since $M\ll N$, $\mathbf{J}$ is inherently rank-deficient. By the Rank-Nullity Theorem, there exists a non-trivial null space $\mathcal{N}(\mathbf{J})$ with dimension $d\geq N-M$. For any perturbation vector $\boldsymbol{\delta}\in\mathcal{N}(\mathbf{J})$, the acoustic observation remains invariant: $\mathcal{W}(\Omega_0+\boldsymbol{\delta})\approx\mathcal{W}(\Omega_0)$. □



(a) SM **(b) Floormap Variations**

Figure 4: Illustration of the influence of the room on RIR. (a): Similarity Matrix (SM) of RIRs with Different Variations of the Room; **(b):** The variations in the floormap

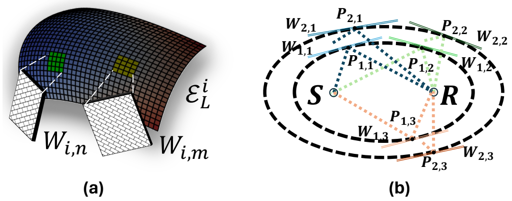


Figure 5: Illustration of the One-To-Many Problem.

Physically, the solution set is a continuous manifold. As shown in Fig. 5, the restriction of one single bounce forms a solution manifold $\mathcal{E}_L^i$. Any discrete point on this manifold represents a mathematically valid reflection candidate. For instance, two distinct candidates, $W_{i,m}$ and $W_{i,n}$, lie on the same manifold $\mathcal{E}_L^i$. Both candidates yield identical acoustic path lengths, yet the geometry is different. Consequently, without external priors, we cannot distinguish these structurally distinct layouts, rendering the naive inference intractable.

To resolve the ambiguity, we adopt a probabilistic perspective. Instead of seeking a deterministic $\mathcal{W}^{-1}$, we formulate the reconstruction task as a conditional generation problem. Specifically, we aim to model a posterior $p(\Omega|S_{\text{obs}})$, where Ω is the geometry and S_{obs} is the set of acoustic observation signals. Using Bayes' theorem, we decompose the intractable posterior into two distinct potentials:

$$p(\Omega|S_{\text{obs}})\propto\underbrace{p(S_{\text{obs}}|\Omega)}_{\text{Physical Likelihood}}\cdot\underbrace{p(\Omega)}_{\text{Geometry Prior}}. \quad (3)$$

This decomposition allows us to address the two fundamental challenges of the task separately:

- **Geometry Prior**: $p(\Omega)$ constrains the solution space to topologically valid floorplans, effectively narrowing down the search within the null space $\mathcal{N}(\mathbf{J})$.

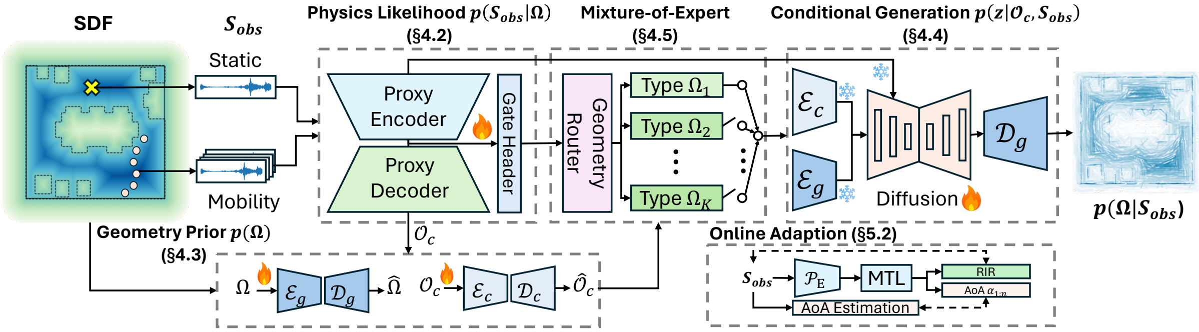


Figure 6: Overview Framework of Loom

- **Physics Likelihood**: $p(S_{\text{obs}}|\Omega)$ ensures the reconstructed geometry satisfies the wave propagation laws governed by $\mathcal{W}$.

Consequently, we transform an underdetermined inverse problem into the modeling of these two complementary distributions. We illustrate the overall framework in Fig. 6.

Observation Signal: A critical challenge is the information sparsity of the input. Currently, the observation S_{obs} is obtained from a single omnidirectional microphone, which collapses the spatial information into a 1D temporal signal. They inherently lack the differential information required to resolve local geometric details. To address this, we draw inspiration from the inherent design of modern smart speakers. Unlike passive recording devices, smart speakers are designed for *active human interaction*. They are typically equipped with microphone arrays to perform Sound Source Localization, determining the user's AoA [85, 86] to optimize beamforming. Concretely, we continuously collect a sequence of observations from routine trajectories:

$$\mathcal{D} = \{S_t\}_{t=1}^{T}, \qquad S_t = \mathcal{W}(\Omega; \mathbf{u}_t) + \eta_t, \tag{4}$$

where $\mathbf{u}_t$ denotes the measurement condition (*e.g.*, relative source position, orientation, *etc.*) at time t. Stacking all measurements yields a composite forward operator $\mathcal{W}_{1:T}(\Omega) = [\mathcal{W}(\Omega; \mathbf{u}_1), \ldots, \mathcal{W}(\Omega; \mathbf{u}_T)]$, whose effective Jacobian has substantially higher rank than that of a single snapshot, thereby shrinking the ambiguity set.

Representation of Floor Geometry: Discrete masks suffer from the vanishing gradient problem in non-boundary regions [15], while vector graphs lack a differentiable topology. To address this, we incorporate the Signed Distance Field (SDF) to represent the floormap, which is defined as

$$\omega(\mathbf{x}) = \begin{cases} -\operatorname{dist}(\mathbf{x}, \partial\Omega), & \mathbf{x} \in \Omega_{\text{in}} \bigcap \Omega_{\text{furn}}^{c} \\ 0, & \mathbf{x} \in \partial\Omega \\ \operatorname{dist}(\mathbf{x}, \partial\Omega), & \mathbf{x} \in \Omega_{\text{out}} \bigcup \Omega_{\text{furn}} \end{cases}, \tag{5}$$

where $\partial\Omega$ denotes the boundary, Ω_{in} denotes the indoor area and Ω_{out} is the outdoor area. Ω_{furn} is the furniture area while Ω_{furn}^{c} is its complementary area. This formulation ensures that the floormap is a continuous, differentiable scalar field.

Coordinate System: In our problem settings, we do not have a global coordinate system under any circumstances. Instead, we adopt a strictly *egocentric* coordinate system. The position of the smart speaker is fixed as the origin $(0, 0, 0)$. Accordingly, all geometric computations are calibrated with respect to the device's local frame of reference. Furthermore, as we do not have the global context, we start the inference of the floormap from a completely empty canvas, with no prior knowledge of the shape, position and orientations.

4.2 Physics Likelihood

We start with the modeling of $p(S_{\text{obs}}|\Omega)$. However, the direct access to $p(S_{\text{obs}}|\Omega)$ is computationally intractable due to complex interactions in high-dimensional spaces, even with ray tracing. Consequently, we introduce a proxy condition O_c and further decompose it as

$$p(S_{\text{obs}}|\Omega) = \int p(S|O_c, \Omega) p(O_c|\Omega) dO_c. \tag{6}$$

Lemma 2. *The estimation of $p(S_{obs}|\Omega)$ is equivalent to the modeling from the signal space to the condition space through a proxy model. Specifically, it requires learning the mapping:*

$$\mathcal{P} : S_{obs} \mapsto O_c$$

Proof. We can rewrite $p(S_{\text{obs}}|O_c, \Omega)$ as $p(S_{\text{obs}}|O_c)$ as we assume the O_c restores the necessary spatial semantics to restore S_{obs}. We reverse it as $p(S_{\text{obs}}|O_c) \propto p(O_c|S_{\text{obs}})/p(O_c)$, where we omit $p(S_{\text{obs}})$ as it is fixed constant. Then we can rewrite Eq. 6 as

$$\begin{aligned} p(S_{\text{obs}}|\Omega) &\approx \int \frac{p(O_c|S_{\text{obs}})}{p(O_c)} p(O_c|\Omega) dO_c \\ &= \mathbb{E}_{O_c \sim p(O_c|\Omega)} \left[\frac{p(O_c|S_{\text{obs}})}{p(O_c)} \right]. \end{aligned} \tag{7}$$

To this end, in order to model the physics likelihood, we are required to learn the projection from the observed measurements to the proxy condition O_c. □

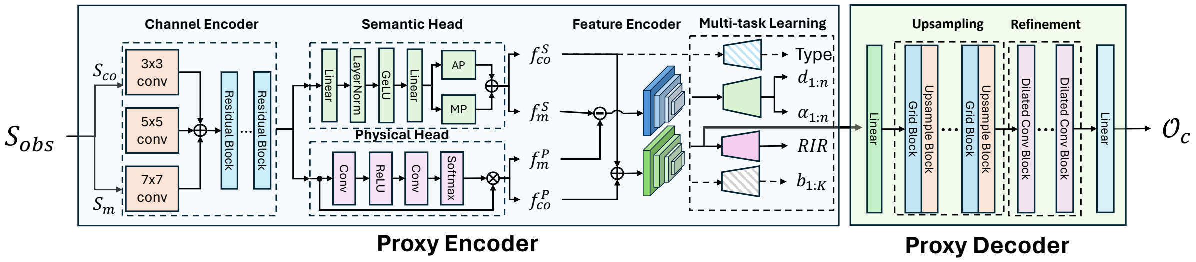


Figure 7: Framework of Proxy Model

Proxy Condition $\mathcal{O}_c$: Intuitively, the proxy condition serves two purposes. First, it should be a low-dimensional intermediary that bridges the raw signals with the fine-grained geometry. Moreover, recall that we start from a completely blank canvas to infer the floormap, the proxy condition should also provide a good initialization of the canvas. Motivated by these, we choose the binary *occupancy map* as the proxy condition $\mathcal{O}_c$, which abstracts essential structural elements and preserves critical physical semantics, including the boundaries, shape, and positions. Furthermore, such an occupancy map holds potential to reduce ambiguity in the inverse problem, acting as the spatial scaffold that can be refined by the subsequent stages via the geometry priors.

Network Architecture: We illustrate the design of the proxy model in Fig. 7. Generally, the model adopts an encoder-decoder structure and accepts dual-branch input, *i.e.*, $S_{\text{obs}}^i = \{S_{\text{co}}^i, S_{\text{m}}^i\}$. Here S_{co}^i means the channel measurements from the co-located smart speakers, while S_{m}^i represents the recordings from the continuous human interactions. The encoder processes these measurements through multi-scale residual blocks to capture temporal-frequency features, followed by the semantic and physical heads. These features are then aggregated via transformers in the feature encoder. The decoder reconstructs the occupancy map $\mathcal{O}_c$.

Multi-Task Learning: Notably, we design the encoder via multi-task learning:

$$\mathcal{P}_E(S_{\text{obs}}) \mapsto (\hat{S}_{\text{co}}, \phi, d, \text{Bounce}, \text{Room Type}). \tag{8}$$

This multi-task design is motivated by two key insights. First, it encourages the encoder to learn richer spatial semantics by leveraging complementary supervisory signals. By jointly optimizing these objectives, the encoder is encouraged to capture spatial cues more effectively. Second, the multi-task framework supports online adaptation to unseen environments, enabling continual self-improvement of the proxy model, as detailed in §5.2.

Proxy Encoder: The acoustic measurements encapsulate various temporal-frequency information. The early reflections denote the high-frequency geometry while the reverberations contain the volume information. Accordingly, we incorporate a multi-scale residual block to capture such information. They are then fed into stacked residual blocks for feature extraction. We then decompose them into semantic head and physical head. The semantic head encodes the global properties while the physical head aims to extract the dynamic information. In the feature encoder, we incorporate the differential network to pinpoint the relative movements. Meanwhile, we add the semantic information to form the complete scene semantics. They are then fed into two transformer encoders for feature extraction.

Proxy Decoder: The decoder reconstructs the high-resolution occupancy map $\mathcal{O}_c \in \mathbb{R}^{H \times W}$. Recovering absolute wall positions from translation-invariant convolutions is challenging. To address this, we introduce a coordinate injection mechanism. At each upsampling stage k, we concatenate the normalized spatial coordinates to the feature map. The features are then processed by residual upsampling blocks to progressively double the resolution. At the final resolution, we employ a dilated refinement module to enforce local geometric continuity. The module consists of stacked convolution layers with increasing dilation rates, which expands the receptive field exponentially. Finally, a prediction head maps the refined features to the pixel-wise occupancy probability.

4.3 Geometry Prior

In this section, we model the geometry prior $p(\Omega)$. It aims to learn topologically valid floorplans, thereby shrinking the search space. However, the space of all possible matrices is vast, hence directly optimizing the geometry Ω in the high-dimensional pixel space is intractable and unstable. Our key observation is that the subset of matrices representing valid floor plans lies in a finite set, including the topologically closed, piecewise-linear walls and the symmetry structures. In other words, such a subset forms a low-dimensional manifold embedded within this space. To explicitly model the geometry prior, we employ a Variational Autoencoder (VAE) to learn a compact latent embedding. Specifically, we learn an encoder $\mathcal{E}_g$ that converts $\Omega_i \in \mathbb{R}^{H \times W}$ to a latent code $z_i \in \mathbb{R}^d$, where $d \ll H \times W$, *i.e.*,

$$q_\phi(\mathbf{z}|\Omega) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\phi(\Omega), \boldsymbol{\sigma}_\phi^2(\Omega)), \tag{9}$$

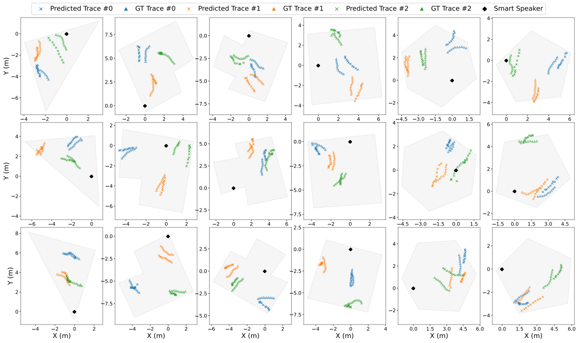


Figure 8: Trajectory Predictions of Proxy Model

where $\mu_\phi(\cdot)$ and $\sigma^2_\phi(\cdot)$ denotes the mean and variance operator, respectively. Meanwhile, we learn a decoder $\mathcal{D}_g$ that recovers the floorplan from any valid latent codes, *i.e.*,

$$p_\psi(\Omega|\mathbf{z}) = \mathcal{N}(\Omega; \mathcal{D}_\psi(\mathbf{z}), \mathbf{I}). \tag{10}$$

Effectively, the decoder learns the geometry prior and produces a structurally plausible floorplan.

4.4 Conditional Generation

Recall Eq. (3), we formulate the floormap inference task as a conditional generation task. In the previous section, we learn a proxy condition, *i.e.*, an occupancy map O_c and model the geometry prior by casting the geometric features into a latent space $\mathbf{z}$, where $\mathbf{z} = \mathcal{E}_g(\Omega)$. Then the remaining step is to model $p(\mathbf{z}|O_c)$, where we leverage the occupancy map O_c as the spatial condition to infer the latent code $\mathbf{z}$. To address this, we rewrite $p(\mathbf{z}|O_c)$ as $p(\mathbf{z}|O_c, S_{\text{obs}})$ and leverage the conditional diffusion model to represent the distribution.

Latent Diffusion Process: Specifically, we view the original latent $\mathbf{z}_0$ as the data point, and inject Gaussian noises as

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{11}$$

Here $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, where $\{\beta_t\}_{t=1}^T$ is a variance schedule. Physically, as $t \longrightarrow T$, the latent code $\mathbf{z}_t$ approaches an isotropic Gaussian distribution, where all structural information of the floorplan is lost. The goal of reconstruction is to reverse this process: recovering the structured geometry $\mathbf{z}_0$ from noise $\mathbf{z}_T$, guided by O_c and S_{obs}, *i.e.*,

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{z}_t, t, O_c, S_{\text{obs}}), \Sigma_\theta(\mathbf{z}_t, t)). \tag{12}$$

Here, the mean $\boldsymbol{\mu}_\theta$ is the core component to be learned. Instead of predicting $\boldsymbol{\mu}_\theta$ directly, we train a neural network $\boldsymbol{\epsilon}_\theta$ to predict the noise $\boldsymbol{\epsilon}$ added in the forward process. The relationship is given by:

$$\boldsymbol{\mu}_\theta = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{z}_t, t, O_c, S_{\text{obs}})\right). \tag{13}$$

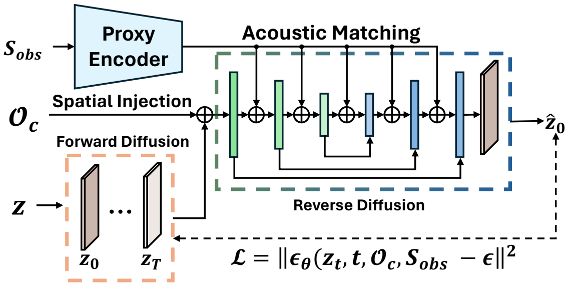


Figure 9: Architecture of PG-LDM

Physics-Guided Optimization: Why does the training of the PG-LDM result in the optimized reconstruction? We analyze this from the score-based generative modeling. Specifically, the optimization of LDM is equivalent to the optimization of the conditional score function,

$$s(z_t, O_c, S_{\text{obs}}) = \log p_t(z_t|O_c, S_{\text{obs}}). \tag{14}$$

Lemma 3. *The optimization of PG-LDM minimizes an energy potential defined by three competing forces: geometry prior, spatial alignment, and acoustic matching, i.e.,*

$$\epsilon^*_\theta \propto \nabla_{z_t}(\log p_t(z_t) + \log p_t(O_c|z_t) + \log p_t(S_{obs}|z_t)) \tag{15}$$

We leave the proof to Appendix C. Therefore, the PG-LDM reconstructs the floorplan by converging toward a state that simultaneously satisfies learned spatial priors, semantic alignment, and physical acoustic consistency.

Network Design: As shown in Fig. 9, we implement the model with a dual-pathway architecture. Since O_c represents the spatial gradient field, we inject O_c via concatenation at the input level. At the same time, we model $\nabla \log p(S_{\text{obs}}|z_t)$ via semantic modulation. We inject the global acoustic scene as a semantic condition, and the network learns to match the structure generated with the priors. We employ a U-Net backbone to integrate these forces and predict the noise $\boldsymbol{\epsilon}$. We optimize the PG-LDM using MSE loss between the predicted noise and the ground truth noise.

4.5 Geometry Expert

Real-world residential environments exhibit vast structural diversity. Encoding the geometric priors of such diverse spaces in a single monolithic model inevitably creates capacity bottlenecks. To address this, Loom introduces a Mixture-of-Experts architecture. At the core of this mechanism is the acoustic proxy network. While its primary task is to invert the acoustic physical likelihood into an explicit geometric condition, it simultaneously employs a classification head to predict the room type. Our key insight is that different room functions inherently imprint distinct macroscopic acoustic signatures. By forcing the network to learn these semantic acoustic signatures explicitly, we obtain a highly reliable layout condition. Once the proxy network processes the acoustic measurements, the predicted *room type* acts as a gating signal. Instead of feeding the geometric conditions into a monolithic generator, the MoE router dynamically dispatches the features to a highly specialized Geometry Expert, which is an independent PG-LDM pre-trained exclusively on a specific spatial category. By compartmentalizing the complicated prior distributions, each expert is freed from the burden of generalizing across contradictory room semantics.

5 IMPLEMENTATION

5.1 Implementation Details

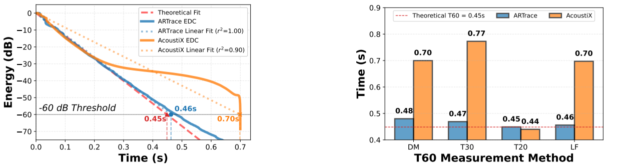
We implement Loom based on Pytorch. The model is trained on 2 NVIDIA H800 80GB and tested on 1 NVIDIA RTX 4090. We list our training details and parameters in Appendix B. We will elaborate on the implementation details below.

Acoustic Data Collection: We use COTS devices [2, 3] for real-world measurements. We leverage log sweep [22] to measure the RIRs, with a frequency band of 20-2kHz. The user continuously moves around and plays the sound using the mobile phone. We sample 10 consecutive points for each trajectory path, which consumes less than 1 minute. For each room, we record three paths at most. We do not limit the paths of the users. Notably, our data collection is significantly less burdensome than prior active scanning methods.

Floorplan Data Preprocessing: Our pipeline incorporates three distinct categories of floorplan data. First, we synthesize a dataset of empty rooms encompassing diverse geometries, including triangular, L-shaped, T-shaped, shoebox, hexagonal, and pentagonal layouts. Second, we leverage the 3D-FRONT dataset [23] to represent complex, furnished real-world environments, benefiting from its rich semantic annotations and diverse room types. Finally, we collect real-world indoor scenes as shown in Fig. 11. We capture point clouds using Polycam [11], process them in Blender, and export them into the Mitsuba XML format [63] for learning.

5.2 Online Adaptation

Deploying Loom into real-world domestic environments inevitably encounters a severe domain shift, including unmodeled hardware frequency responses, background noise, and hardware imperfections. Crucially, in practical deployments, we have absolutely zero access to ground-truth room layouts to fine-tune the model. To bridge this gap, Loom incorporates an online adaptation mechanism to calibrate the model. Our core insight is that while the geometric layout is hidden, intermediate acoustic properties can be directly measured. First, the RIR can be directly measured from the playback of the smart speaker. Second, modern smart speakers can reliably track the spatial direction of ambient sound sources. As shown in Fig. 6, we leverage the measured RIR as the supervision signal and estimate the user's continuous AoA trajectory as the pseudo-labels. During online deployment, gradients derived with these physical constraints are back-propagated to fine-tune the proxy encoder.



(a) EDC **(b) Different Measurements**

Figure 10: Comparisons of acoustic physics simulations

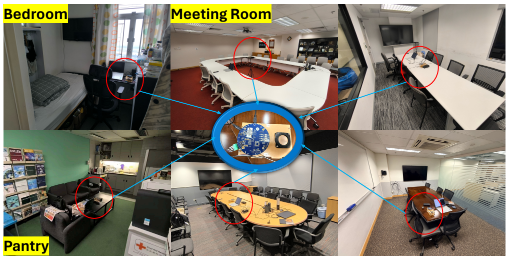


Figure 11: Experiment Settings

5.3 Data Engine

Training data-hungry neural models requires a highly diverse dataset of paired acoustic responses and accurate floorplans. Existing differentiable simulators fall short in key respects: they either lack differentiability, cannot import custom scenes [73, 74], or compromise core acoustic propagation physics [32, 45]. For example, recent EM-adapted differentiable ray tracers (*e.g.*, AcoustiX) rely on heavy heuristics to approximate sound propagation, including applying post-hoc frequency kernels, and randomly perturbing phase coherence. As shown in Fig. 10, these approximations introduce substantial macroscopic distortions: AcoustiX produces a warped Energy Decay Curve (EDC) and systematically overestimates the standard reverberation time (T60), failing to provide the high-fidelity acoustic priors required by Loom. To address these limitations, we develop ARTrace, an acoustic-specific differentiable ray tracer built on OptiX [67] and DrJit [33]. We defer implementation details to Appendix A. As validated in Fig. 10, ARTrace closely matches the theoretical thermodynamic EDC trajectory and yields T60 values consistent with standardized measurements.

6 EVALUATION

6.1 Comparative Study

Metrics: We evaluate all methods in both in-scene and cross-scene settings, measuring reconstruction quality and generalization to unseen scenes. We report SSIM, IoU, and Chamfer Distance for geometric accuracy.

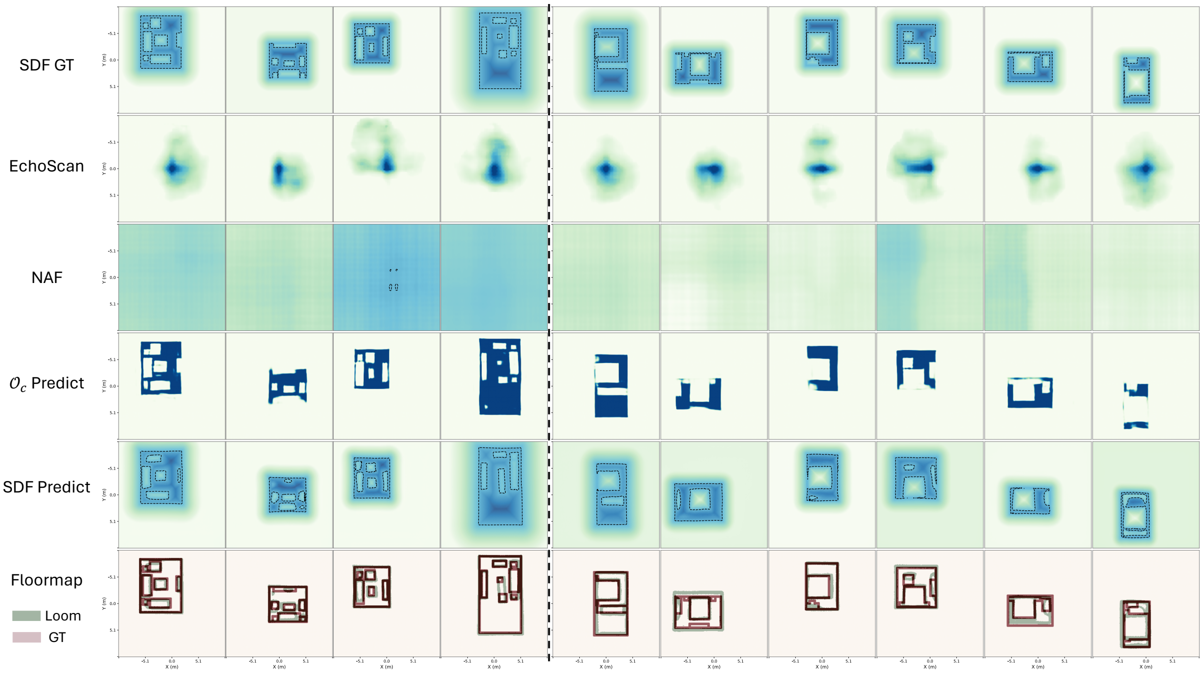


Figure 12: Visualizations of Inferring Furnished Room. Left: In-Scene Inference **Right:** Cross-Scene Inference

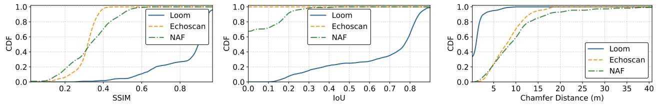
Baseline: We compare Loom with two representative baselines. EchoScan [96] is a state-of-the-art echo-based method that predicts wall boundaries from acoustic reflections. NAF [52] is an acoustic NeRF-style framework that learns an implicit field from acoustic observations. For NAF, we use the predicted trajectories in Fig. 8 to provide the required local coordinates. We first train a NAF model per scene, then fit a shared linear probe that maps the learned embeddings to the ground-truth SDF values across scenes.

Performances of Furnished Rooms: As shown in Fig. 12, we visualize our results of furnished rooms. Intuitively, Loom outputs semantic-rich occupancy maps and highly accurate SDF predictions with corresponding floormaps. Conversely, EchoScan and NAF fail to decode meaningful spatial semantics, leading to blurred and noisy predictions. We benchmark our results in Fig. 13 and Fig. 14. For in-scene inference, Loom achieves an average SSIM of 0.8256, outperforming EchoScan with 0.3065 and NAF with 0.3273, corresponding to relative improvements of 169.3% and 152.3%, respectively. The IoU of Loom reaches 0.6504, while EchoScan and NAF obtain near-zero overlap with the ground truth. Moreover, Loom attains a mean Chamfer Distance of 1.2484 m, yielding 6.4x and 7.8x lower error, respectively. This validates Loom's superior ability to infer the room geometry given acoustic measurements. For cross-scene generalization, Loom remains robust and achieves an average SSIM of 0.7700. Meanwhile, Loom attains a mean Chamfer Distance of 1.1703 m, dramatically lower than 6.9958 m for EchoScan and 8.8848 m for NAF. Therefore, Loom transfers effectively to unseen scenes, delivering substantially more accurate and scalable reconstructions than prior methods.

Performances of Empty Rooms: We further evaluate Loom in empty-room settings with diverse room shapes. As shown in Fig. 15, Loom successfully reconstructs rooms with a wide range of shapes. As in Fig. 16, Loom achieves an SSIM of 0.9056 and reaches an IoU of 0.9753, compared with 0.9145 for both baselines. Loom also substantially reduces geometric error, yielding a Chamfer Distance of 2.2116 m, versus 9.4667 m for EchoScan and 9.8516 m for NAF.

Performances of Proxy Models: We also evaluate the performances of our proxy models. As shown in Fig. 12 and Fig. 15, O_c is aligned with the ground truth. However, we also observe blurry boundaries. We attribute this to the regression nature of the proxy model, which tends to produce averaged predictions and does not enforce strong geometric priors. As shown in Fig. 17, without the PG-LDM, the proxy model achieves an SSIM of 0.3035 and an IoU of 0.2517, and attains a Chamfer Distance of 1.3127 m. This behavior is consistent with our design goal. The proxy model captures acoustic propagation cues and provides a coarse spatial hint through O_c, while the PG-LDM injects geometry priors to sharpen boundaries and recover plausible floorplan structure.

Different Room Functionalities: We analyze the performance of Loom across different room functions, including living rooms, bedrooms, and dining rooms. As shown in Fig. 18, bedrooms and dining rooms yield consistently higher



(a) SSIM (b) IoU (c) Chamfer

Figure 13: In-Scene Performances

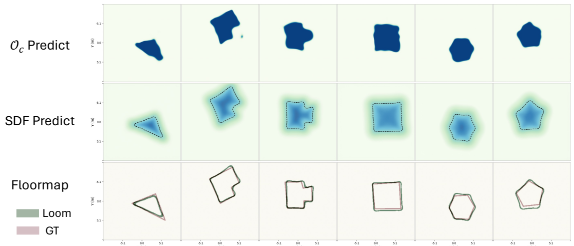


Figure 15: Visualizations of Inferring Empty Rooms

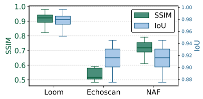


Figure 16: Empty Rooms

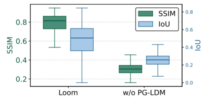


Figure 17: Proxy-Only Result

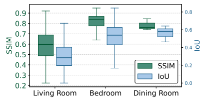


Figure 18: Different Rooms

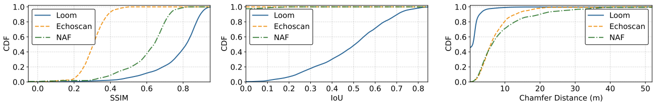
reconstruction fidelity, with SSIM of 0.8163 and 0.7824, respectively. In contrast, living rooms are more challenging, with an SSIM of 0.5860. These results validate the usage of our MoE framework across different room types, and suggest that Loom generalizes well across room types, while larger and more open spaces remain the most difficult due to increased layout complexity.

6.2 MicroBenchmarks

Different Smart Speaker Heights: We evaluate Loom under different speaker heights to examine its robustness to placement. As in Fig. 19, Loom maintains consistently strong performance across all tested heights from 1.1 m to 1.9 m, with SSIM ranging from 0.8956 to 0.9198. Overall, these results suggest that Loom is not sensitive to speaker heights.

Different User Heights: We evaluate Loom under different user heights ranging from 1.1 m to 1.9 m. As shown in Fig. 20, Loom achieves average SSIM consistently above 0.90 and mean occupancy IoU above 0.97. These results indicate that Loom is robust to user-height variations and can reliably infer room layouts under diverse user conditions.

Different Noises: We evaluate Loom under different noise levels. As shown in Fig. 21, increasing noise leads to a gradual degradation in structural similarity, with mean SSIM dropping from 0.8036 at -40 dB to 0.7430 at -25 dB. The mean IoU stays above 0.92 across all settings. Overall, these results indicate that Loom is robust to substantial background noise, preserving reliable geometric predictions.



(a) SSIM (b) IoU (c) Chamfer

Figure 14: Cross-Scene Performances

Different Room Areas: We also study the impact of room areas on Loom. As shown in Fig. 22, Loom remains stable across all tested room sizes. SSIM stays high, ranging from 0.8734 to 0.9247. Overall, Loom generalizes well to different room scales, which we attribute to the accurate acoustic physics prior modeling in our pipeline.

Different Number of Mobile Samples: We evaluate Loom with different numbers of measurements. As shown in Fig. 24, performance improves as more samples are available. When reducing the budget to 10 samples, the degradation is marginal. Overall, these results suggest that Loom can effectively leverage additional measurements, while maintaining strong performance even with a small number of samples.

Different Room Shapes: Fig. 23 shows the CDFs of SSIM and IoU in empty rooms with different shapes. Triangle and pentagonal rooms perform best, reaching a mean SSIM of 0.9333 and 0.9190 and a mean IoU of 0.9842 and 0.9797, respectively. This demonstrates our usage of MoE that routes different room shapes to different geometry priors.

Latency: We benchmark the latency of Loom. The proxy encoder achieves an average latency of 5.9ms per input S_{obs}, while the proxy decoder takes 7.43ms. For the PG-LDM, the end-to-end latency is larger, reaching 439.79ms. Nevertheless, since Loom achieves passive convergence through iterative user interactions, this latency remains acceptable in practice.

6.3 Loom in the Wild

Since real-world rooms are heterogeneous and measurements are quite noisy, we present an in-the-wild study to investigate the real-world practicability of Loom.

Online Adaptation Performance: In real deployments, ground-truth room geometry is unavailable, making supervised fine-tuning infeasible. Instead, we exploit AoA and RIR, to adapt the model to the current environment, as described in §5.2. As shown in Fig. 25, online adaptation substantially improves real-world performance. With the initial inference, the model recovers the room shape coarsely but fails to align it correctly in the local coordinate system. After several epochs of adaptation, the AoA estimates are calibrated, enabling Loom to infer the room geometry at the correct location. This trend is also reflected quantitatively. SSIM increases from 0.36 to 0.44, while the Chamfer Distance drops sharply from 12.2 m to 1.2 m. These results validate the

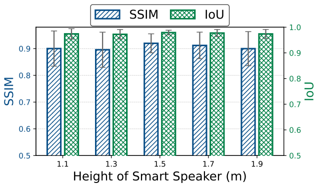


Figure 19: Speaker Heights

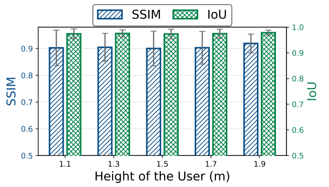


Figure 20: User Heights

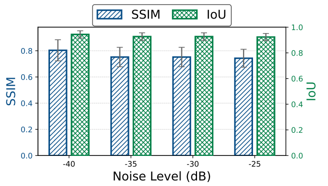


Figure 21: Different Noises

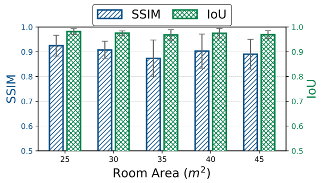
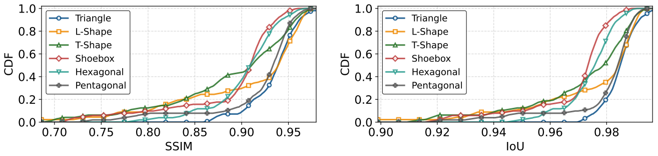


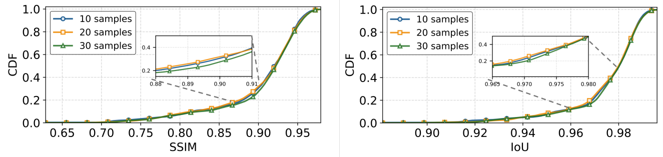
Figure 22: Different Areas



(a) SSIM

(b) IoU

Figure 23: Different Room Shapes



(a) SSIM

(b) IoU

Figure 24: Different Number of Samples

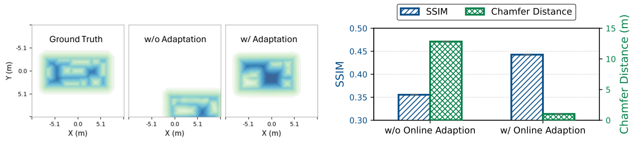


Figure 25: Influence of Online Adaptation

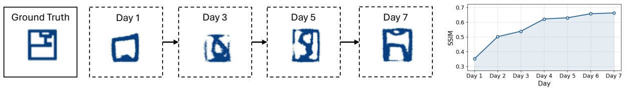


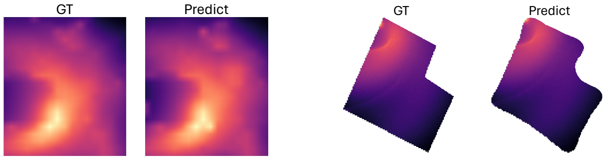
Figure 26: One-Week Iterative Study

effectiveness of our online adaptation module and demonstrate reliable zero-shot calibration in real environments.

Iterative Study: To assess stability over time, we conduct an iterative study spanning one week. During this period, we continuously and opportunistically reuse interaction recordings for online adaptation and inference. As shown in Fig. 26, SSIM improves from 0.35 to 0.66, corresponding to an 88.5% relative gain, and the predicted floorplan becomes progressively more refined. Despite this steady improvement, residual discrepancies from the ground truth remain. We attribute them to two factors. First, real rooms are not strictly static. Everyday human activities move or deform small objects such as comforters and books. These objects are not explicitly modeled in our system, yet they can still affect acoustic propagation. Second, AoA estimates in real deployments are noisy. Using these signals as adaptation targets can introduce bias and lead to cumulative errors over long-term updates. Overall, these in-the-wild results show that Loom can operate reliably under realistic noise and domain shifts.

7 RELATED WORK

7.1 Modeling the Signal Field



(a) Furnished Room

(b) Empty Room

Figure 27: Spatial Distribution of the Acoustic Field

Existing approaches generally fall into four categories, each with distinct limitations. ❶ **Neural Field Rendering:** Since NeRF [59], numerous works have attempted to model wireless [4, 48, 51, 90, 100] or acoustic environments [45, 52, 77] via neural field rendering. These methods represent the environment as a continuous function, mapping coordinates to signal observations. However, the geometry is entangled within the network parameters, failing to output interpretable floorplans. Furthermore, these methods are inherently scene-specific, and require dense, active sampling to fit a single room with precise global coordinates. While recent attempts like NAF [52] try to decode explicit geometry by adding auxiliary layers to the implicit latent, they struggle to generalize across different rooms. EchoNeRF [4] directly treats the occupancy grids as the signal field, rendering it incapable of discovering the geometry of unknown environments from acoustic signals alone. ❷ **Differentiable Ray Tracing (DRT):** DRT methods [11, 30, 32, 35, 36, 45, 64] integrate the physics of wave propagation directly into the optimization loop. However, these methods effectively solve a parameter refinement problem rather than a topology discovery problem. They excel at fine-tuning the material properties [30, 32, 35, 45] or exact positions of known walls [36] but struggle to generate a floorplan from scratch. Without a precise initial guess, which is unavailable in practical setups, DRT methods are prone to getting stuck in local minima, unable to "grow" walls where none were initialized. ❸ **Signal Generative Models:** A parallel line of research leverages generative models to synthesize radio [13, 14] or acoustic maps [54, 60, 71]. They utilize conditional inputs, such as visual images [12, 44] or 3D point clouds [66], to predict signal distributions. These works focus on signal synthesis rather than geometric inference. They do not support the inverse generation of layouts conditioned strictly on sparse, passive signal measurements. ❹ **Acoustic Ranging and Localization:** Other lines of work focus on the ranging and localization [6, 9, 24, 47, 57, 75, 82, 83, 99]. They analyze reflections to track device motion or pinpoint sound sources, where they typically treat multipath reflections as nuisance parameters to be filtered, or strictly as constraints to localize a specific point source. While some works [5, 19, 27, 76, 95] attempt to estimate wall distances, they often rely on simplified signal models, require large microphone arrays [16] or assume simplified rooms without any furniture [69, 82, 96]. These constraints render them impractical for recovering complex residential floorplans using commodity hardware.

7.2 Modeling the Room Geometry

We review existing approaches in two primary categories: ❶ **Visual Methods:** Vision-based approaches represent the floormap as rasterized segmentation masks [46, 62], vectorized planar graphs [8], or geometric primitives (*e.g.*, corners and edges) [49, 50]. However, these models rely heavily on dense, semantic-rich inputs (*e.g.*, RGB images [34, 81], point clouds [87], or user-drawn sketches [79]. They cannot be applied directly to acoustics, as our measurements lack explicit vision-like spatial features. ❷ **Non-Visual Methods:** Recent advances infer floormap from numerous non-visual sources, including acoustics [58, 96, 102], RF radars [10, 103], and user trajectories [61]. However, these approaches face two fundamental barriers to ubiquitous deployment. First, they predominantly rely on active scanning. SLAM-based systems [17, 21, 39, 42, 88, 89] leverage moving agents to triangulate landmarks. Crowdsourcing systems [25, 26, 68, 92, 94, 102] require users to walk along room boundaries to construct a synthetic aperture or trace the room shape. These prevent their large-scale, automated deployment in residential settings. Second, works using RF radars [38, 97, 103] are constrained by hardware directionality and specular reflections. Unlike omnidirectional smart speakers, commercial radars suffer from a limited Field-of-View due to directional beamforming. Furthermore, specialized mmWave hardware lacks the ubiquity of smart speakers, failing to serve as foundational infrastructure for home sensing.

8 DISCUSSIONS AND FUTURE WORK

Privacy: While Loom opportunistically reuses the recordings of in-situ smart speakers, such devices have already become deeply entrenched as ubiquitous domestic infrastructure, indicating a practical acceptance of this acoustic presence. More importantly, compared to vision-based mapping, our acoustic-only paradigm poses significantly lower privacy risks. Moving forward, we plan to develop lightweight algorithms that run entirely on-device.

Home-Level Reconstruction: As a pioneer work, Loom is for the room-level floormap reconstruction. We envision two opportunities for expanding Loom to home-level solution. First, as increasing families own multiple acoustic devices at home [18], future systems could orchestrate a distributed acoustic mesh, collaboratively stitching local room geometries into a global layout. Second, by augmenting acoustics with ubiquitous RF signals like Wi-Fi, a joint system could resolve inter-room topologies and seamlessly extend spatial intelligence across the entire home.

Ambient Sound Sources: Real-world domestic environments are continuously bathed in uncooperative ambient sounds. Future work can exploit foundational audio models to isolate and disentangle these diverse background events and transform these noises into free spatial beacons.

Combination with DRT: Recently, increasing works leverage DRT for building the digital twins [11, 44]. However, they are primarily confined to optimizing local acoustic parameters (*e.g.*, materials). Loom provides the critical structural "0-to-1" initialization required to unlock DRT's full potential in unseen environments. Future systems could tightly couple Loom's generative priors with DRT-based inverse rendering, enabling a closed-loop joint optimization of both fine-grained geometric and material properties.

Signal Field Synthesis: Existing methods leverage ray tracing [55] or neural rendering [45] to reconstruct the signal field. Loom provides a new paradigm without exhaustively traversing the space. As shown in Fig. 27, it enables high-fidelity signal field rendering from in-situ observations. Future work can explore the advanced usage of Loom.

3D Floorplan: Due to the limited size of the COTS smart speakers, the elevation is inherently ambiguous. To recover the 3D floorplan, future work can resort to multi-modality methods that provides richer spatial semantics.

9 CONCLUSION

In this paper, we present Loom, the first-of-its-kind system that repurposes in-situ COTS smart speakers into a scalable neural floorplan inference solution. At the heart of Loom is we formulate this ill-posed inverse problem into a physics-guided conditional generation problem. We design a dedicated proxy network and PG-LDM to represent the acoustic physics likelihood and geometry prior. Meanwhile, we develop systematic methods to boost the reliability under unseen environments. We achieve an average IoU of 0.65 in furnished rooms. We envision Loom will open up many new directions to boost the spatial intelligence.

REFERENCES

[1] 2026. Fresnel equations. https://en.wikipedia.org/wiki/Fresnel_equations. Accessed: 2026-03-09.

[2] 2026. Speakers & Receivers | AS05308AS-R – PUI Audio. https://puiaudio.com/product/speakers-and-receivers/as05308as-r. Accessed: 2026-03.

[3] 2026. USB Audio Streaming: UMA-8-SP USB mic array. https://www.minidsp.com/products/usb-audio-interface/uma-8-sp-detail. Accessed: 2026-03.

[4] Chaitanya Amballa, Sattwik Basu, Yu-Lin Wei, Zhijian Yang, Mehmet Ergezer, and Romit Roy Choudhury. 2025. Can NeRFs See without Cameras? *arXiv preprint arXiv:2505.22441* (2025).

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

[6] Guanyu Cai and Jiliang Wang. 2024. Locating your smart devices with a single speaker. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. 28–40.

[7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14455–14465.

[8] Jiacheng Chen, Chen Liu, Jiaye Wu, and Yasutaka Furukawa. 2019. Floor-sp: Inverse cad for floorplans by sequential room-wise shortest path. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2661–2670.

[9] Sheng Chen, Renrui Tan, Zhiming Wang, Xinyu Tong, and Keqiu Li. 2023. VoiceMap: Autonomous mapping of microphone array for voice localization. *IEEE Internet of Things Journal* 11, 2 (2023), 2909–2923.

[10] Weiyan Chen, Hongliu Yang, Xiaoyang Bi, Rong Zheng, Fusang Zhang, Peng Bao, Zhaoxin Chang, Xujun Ma, and Daqing Zhang. 2023. Environment-aware multi-person tracking in indoor environments with mmWave radars. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–29.

[11] Xingyu Chen, Jianrong Ding, Kai Zheng, Xinmin Fang, Xinyu Zhang, Chris Xiaoxuan Lu, and Zhengxiong Li. 2025. InverTwin: Solving Inverse Problems via Differentiable Radio Frequency Digital Twin. *arXiv preprint arXiv:2508.14204* (2025).

[12] Xingyu Chen, Zihao Feng, Ke Sun, Kun Qian, and Xinyu Zhang. 2024. Rfcanvas: Modeling rf channel by fusing visual priors and few-shot rf measurements. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. 464–477.

[13] Xingyu Chen and Xinyu Zhang. 2023. Rf genesis: Zero-shot generalization of mmwave sensing through simulation-based data synthesis and generative diffusion models. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 28–42.

[14] Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han. 2024. RF-diffusion: Radio signal generation via time-frequency diffusion. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 77–92.

[15] Emanuele De Leo. 2025. *Depth Estimation from technical drawings and 3D Mesh Reconstruction with Deep Learning*. Ph. D. Dissertation. Politecnico di Torino.

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

[17] Erqun Dong, Jingao Xu, Chenshu Wu, Yunhao Liu, and Zheng Yang. 2019. Pair-navi: Peer-to-peer indoor navigation with mobile visual slam. In *IEEE INFOCOM 2019-IEEE conference on computer communications*. IEEE, 1189–1197.

[18] Edison Research. 2024. The Infinite Dial 2024. https://www.edisonresearch.com/the-infinite-dial-2024/ Reports that 46% of the U.S. population aged 12+ owns a smart speaker.

[19] Youssef El Baba, Andreas Walther, and Emanuël AP Habets. 2017. 3D room geometry inference based on room impulse response stacks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 5 (2017), 857–872.

[20] Moustafa Elhamshary, Moustafa Alzantot, and Moustafa Youssef. 2018. JustWalk: A crowdsourcing approach for the automatic construction of indoor floorplans. *IEEE Transactions on Mobile Computing* 18, 10 (2018), 2358–2371.

[21] Christine Evers and Patrick A Naylor. 2018. Acoustic slam. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 9 (2018), 1484–1498.

[22] Angelo Farina. 2000. Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique. *Journal of The Audio Engineering Society* (2000). https://api.semanticscholar.org/CorpusID:9614437

[23] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 2021. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10933–10942.

[24] Yongjian Fu, Yongzhao Zhang, Hao Pan, Yu Lu, Xinyi Li, Lili Chen, Ju Ren, Xiong Li, Xiaosong Zhang, and Yaoxue Zhang. 2024. Pushing the Limits of Acoustic Spatial Perception via Incident Angle Encoding. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 52 (May 2024), 28 pages. https://doi.org/10.1145/3659583

[25] Ruipeng Gao, Mingmin Zhao, Tao Ye, Fan Ye, Guojie Luo, Yizhou Wang, Kaigui Bian, Tao Wang, and Xiaoming Li. 2016. Multi-story indoor floor plan reconstruction via mobile crowdsensing. *IEEE Transactions on Mobile Computing* 15, 6 (2016), 1427–1442.

[26] Ruipeng Gao, Mingmin Zhao, Tao Ye, Fan Ye, Yizhou Wang, Kaigui Bian, Tao Wang, and Xiaoming Li. 2014. Jigsaw: Indoor floor plan reconstruction via mobile crowdsensing. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. 249–260.

[27] Shan Gao, Xihong Wu, and Tianshu Qu. 2022. Room geometry blind inference based on the localization of real sound source and first order reflections. *arXiv preprint arXiv:2207.10478* (2022).

[28] Grand View Research, Inc. 2023. Smart Speaker Market Size To Reach $50.19 Billion By 2030. Grand View Research press release. https://www.grandviewresearch.com/press-release/global-smart-speakers-market Accessed: 13 March 2026.

[29] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. 2021. Embodied intelligence via learning and evolution. *Nature communications* 12, 1 (2021), 5721.

[30] Xueqiang Han, Tianyue Zheng, Tony Xiao Han, and Jun Luo. 2025. RayLoc: Wireless indoor localization via fully differentiable ray-tracing. *arXiv preprint arXiv:2501.17881* (2025).

[31] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[32] Jakob Hoydis, Fayçal Aït Aoudia, Sebastian Cammerer, Merlin Nimier-David, Nikolaus Binder, Guillermo Marcus, and Alexander Keller. 2023. Sionna RT: Differentiable ray tracing for radio propagation modeling. In *2023 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 317–321.

[33] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, and Delio Vicini. 2022. Dr. jit: A just-in-time compiler for differentiable rendering. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–19.

[34] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. 2022. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1654–1663.

[35] Xutong Jin, Chenxi Xu, Ruohan Gao, Jiajun Wu, Guoping Wang, and Sheng Li. 2024. DiffSound: Differentiable modal sound rendering and inverse rendering for diverse inference tasks. In *ACM SIGGRAPH 2024 Conference Papers*. 1–12.

[36] Tobias Jüterbock, Ugo Finnendahl, Markus Worchel, Daniel Wujecki, Marc Alexa, and Stefan Weinzierl. 2025. misuka: An open-source differentiable room acoustic renderer. In *Proceedings of Meetings on Acoustics*, Vol. 58. Acoustical Society of America, 022004.

[37] Sukanth Kalivarathan, Muhmmad Abrar Raja Mohamed, Aswathy Ravikumar, and S Harini. 2025. Intelligence of things: A spatial context-aware control system for smart devices. *arXiv preprint arXiv:2504.13942* (2025).

[38] Usman Mahmood Khan, Raghav H Venkatnarayan, and Muhammad Shahzad. 2020. RFMap: Generating indoor maps using RF signals. In *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 133–144.

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

[40] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79 – 86. https://doi.org/10.1214/aoms/1177729694

[41] Heinrich Kuttruff. 2016. *Room Acoustics* (6 ed.). CRC Press, Boca Raton, FL. https://books.google.com/books?id=JW4NDgAAQBAJ

[42] Haowen Lai, Zhiwei Zheng, and Mingmin Zhao. 2025. RF-Based 3D SLAM Rivaling Vision Approaches. In *ACM International Conference on Mobile Computing and Networking (MobiCom)*.

[43] John Lambert, Yuguang Li, Ivaylo Boyadzhiev, Lambert Wixson, Manjunath Narayana, Will Hutchcroft, James Hays, Frank Dellaert, and Sing Bing Kang. 2022. Salve: Semantic alignment verification for floorplan reconstruction from sparse panoramas. In *European Conference on Computer Vision*. Springer, 647–664.

[44] Zitong Lan, Yiwei Tang, Yuhan Wang, Haowen Lai, Yiduo Hao, and Mingmin Zhao. 2025. Building Audio-Visual Digital Twins with Smartphones. *arXiv preprint arXiv:2512.10778* (2025).

[45] Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. 2024. Acoustic volume rendering for neural impulse response fields. *Advances in Neural Information Processing Systems* 37 (2024), 44600–44623.

[46] Chen Liu, Jiaye Wu, and Yasutaka Furukawa. 2018. Floornet: A unified framework for floorplan reconstruction from 3d scans. In *Proceedings of the European conference on computer vision (ECCV)*. 201–217.

[47] Manni Liu and Zhichao Cao. 2024. SoundFlower: A robust sound source localization system for voice assistants. In *2024 IEEE 30th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 92–99.

[48] Mufan Liu, Cixiao Zhang, Qi Yang, Yujie Cao, Yiling Xu, Yin Xu, Shu Sun, Mingzeng Dai, and Yunfeng Guan. 2025. Rasterizing Wireless Radiance Field via Deformable 2D Gaussian Splatting. *arXiv preprint arXiv:2506.12787* (2025).

[49] Yiyi Liu, Chunyang Liu, Bohan Wang, Weiqin Jiao, Bojian Wu, Lubin Fan, Yuwei Chen, Fashuai Li, and Biao Xiong. 2025. CAGE: Continuity-Aware edGE Network Unlocks Robust Floorplan Reconstruction. *arXiv preprint arXiv:2509.15459* (2025).

[50] Yuzhou Liu, Lingjie Zhu, Xiaodong Ma, Hanqiao Ye, Xiang Gao, Xianwei Zheng, and Shuhan Shen. 2024. PolyRoom: Room-aware transformer for floorplan reconstruction. In *European Conference on Computer Vision*. Springer, 322–339.

[51] Haofan Lu, Christopher Vattheuer, Baharan Mirzasoleiman, and Omid Abari. 2024. Newrf: A deep learning framework for wireless radiation field reconstruction and channel prediction. *arXiv preprint arXiv:2403.03241* (2024).

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

[53] Sheng Lyu and Chenshu Wu. 2025. ASE: Practical Acoustic Speed Estimation Beyond Doppler via Sound Diffusion Field. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 115 (Sept. 2025), 26 pages. https://doi.org/10.1145/3749475

[54] Sheng Lyu, Yuemin Yu, and Chenshu Wu. 2025. Temporal Modeling of Room Impulse Response Generation via Multi-Scale Autoregressive Learning. In *Proc. Interspeech 2025*. 923–927.

[55] Ruichun Ma, Shicheng Zheng, Hao Pan, Lili Qiu, Xingyu Chen, Liangyu Liu, Yihong Liu, Wenjun Hu, and Ju Ren. 2024. Automs: Automated service for mmwave coverage optimization using low-cost metasurfaces. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 62–76.

[56] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 2025. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6924–6934.

[57] Wenguang Mao, Wei Sun, Mei Wang, and Lili Qiu. 2020. DeepRange: Acoustic ranging via deep learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–23.

[58] Chuize Meng, Shan Jiang, Mengning Wu, Xuan Xiao, Dan Tao, and Ruipeng Gao. 2022. BatMapper-Plus: Smartphone-Based Multi-level Indoor Floor Plan Construction via Acoustic Ranging and Inertial Sensing. In *International Conference on Wireless Algorithms, Systems, and Applications*. Springer, 155–167.

[59] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[60] Federico Miotello, Luca Comanducci, Mirco Pezzoli, Alberto Bernardini, Fabio Antonacci, and Augusto Sarti. 2024. Reconstruction of sound field through diffusion models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1476–1480.

[61] Claudio Mura, Renato Pajarola, Konrad Schindler, and Niloy Mitra. 2021. Walk2map: Extracting floor plans from indoor walk trajectories. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 375–388.

[62] Nelson Nauata, Kai-Hung Chang, Chin-Yi Cheng, Greg Mori, and Yasutaka Furukawa. 2020. House-gan: Relational generative adversarial networks for graph-constrained house layout generation. In *European Conference on Computer Vision*. Springer, 162–177.

[63] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. 2019. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (ToG)* 38, 6 (2019), 1–17.

[64] Tribhuvanesh Orekondy, Pratik Kumar, Shreya Kadambi, Hao Ye, Joseph Soriaga, and Arash Behboodi. 2023. Winert: Towards neural ray tracing for wireless channel modelling and differentiable simulations. In *The Eleventh International Conference on Learning*

Representations.

[65] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. 2025. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805* (2025).

[66] Kyoungjun Park, Yifan Yang, Changhan Ge, Lili Qiu, and Shiqi Jiang. 2025. Diffusionˆ 2: Turning 3D Environments into Radio Frequency Heatmaps. *arXiv preprint arXiv:2510.02274* (2025).

[67] Steven G Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, et al. 2010. Optix: a general purpose ray tracing engine. *Acm transactions on graphics (tog)* 29, 4 (2010), 1–13.

[68] Swadhin Pradhan, Ghufran Baig, Wenguang Mao, Lili Qiu, Guohai Chen, and Bo Yang. 2018. Smartphone-based acoustic indoor space mapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–26.

[69] Tilak Rajapaksha, Xiaojun Qiu, Eva Cheng, and Ian Burnett. 2016. Geometrical room geometry estimation from room impulse responses. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 331–335.

[70] Alexander Rath, Pascal Grittmann, Sebastian Herholz, Philippe Weier, and Philipp Slusallek. 2022. EARS: efficiency-aware russian roulette and splitting. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–14.

[71] Anton Ratnarajah, Shi-Xiong Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. 2022. FAST-RIR: Fast neural diffuse room impulse response generator. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 571–575.

[72] SAC INSiGHT Private Limited. 2025. Top 10 Companies in Smart Home Industry Shaping Connected Living in 2025. LinkedIn Pulse. https://www.linkedin.com/pulse/top-10-companies-smart-home-industry-shaping-connected-living-uwjqf/ Accessed: 13 March 2026.

[73] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9339–9347.

[74] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. 2018. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 351–355.

[75] Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury. 2020. Voice localization using nearby wall reflections. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[76] Oliver Shih and Anthony Rowe. 2019. Can a phone hear the shape of a room?. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*. 277–288.

[77] Chen Si, Qianyi Wu, Chaitanya Amballa, and Romit Roy Choudhury. 2025. Explicit Context-Driven Neural Acoustic Modeling for High-Fidelity RIR Generation. *arXiv preprint arXiv:2509.15210* (2025).

[78] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[79] Muhammad Usman, Abdullah Almulhim, Mohammad Alaseri, Mona Alzahran, Hamzah Luqman, and Saeed Anwar. 2024. Sketch–to–3D: Transforming Hand-Sketched Floorplans into 3D Layouts. In *2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 359–366.

[80] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

[81] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5294–5306.

[82] Mei Wang, Wei Sun, and Lili Qiu. 2021. {MAVL}: Multiresolution analysis of voice localization. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*. 845–858.

[83] Weiguo Wang, Jinming Li, Yuan He, and Yunhao Liu. 2020. Symphony: Localizing multiple acoustic sources with a single microphone array. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 82–94.

[84] Yiwen Wang, Albert Kai-Sun Wong, S-H Gary Chan, and Wai Ho Mow. 2023. Leto: Crowdsourced radio map construction with learned topology and a few landmarks. *IEEE Transactions on Mobile Computing* 23, 4 (2023), 2795–2812.

[85] Yu-Lin Wei and Romit Roy Choudhury. 2021. Estimating angle of arrival (aoa) of multiple echoes in a steering vector space. *arXiv preprint arXiv:2109.13072* (2021).

[86] Yu-Lin Wei, Rui Li, Abhinav Mehrotra, Romit Roy Choudhury, and Nic Lane. 2021. Inferring facing direction from voice signals. *arXiv preprint arXiv:2109.13094* (2021).

[87] Honghao Xu, Juzhan Xu, Zeyu Huang, Pengfei Xu, Hui Huang, and Ruizhen Hu. 2024. Fri-net: Floorplan reconstruction via room-wise implicit representation. In *European Conference on Computer Vision*. Springer, 1–17.

[88] Jingao Xu, Hao Cao, Danyang Li, Kehong Huang, Chen Qian, Longfei Shangguan, and Zheng Yang. 2020. Edge Assisted Mobile Semantic Visual SLAM. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*. 1828–1837. https://doi.org/10.1109/INFOCOM41043.2020.9155438

[89] Jingao Xu, Hao Cao, Zheng Yang, Longfei Shangguan, Jialin Zhang, Xiaowu He, and Yunhao Liu. 2022. SwarmMap: Scaling Up Real-time Collaborative Visual SLAM at the Edge. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*. USENIX Association, Renton, WA, 977–993. https://www.usenix.org/conference/nsdi22/presentation/xu

[90] Kang Yang, Yuning Chen, and Wan Du. 2025. Gwrf: A generalizable wireless radiance field for wireless signal propagation modeling. *arXiv e-prints* (2025), arXiv–2502.

[91] Zhijian Yang and Romit Roy Choudhury. 2023. MapLearn: Indoor mapping using audio. (2023).

[92] Zheng Yang, Chenshu Wu, and Yunhao Liu. 2012. Locating in fingerprint space: Wireless indoor localization with little human intervention. In *Proceedings of the 18th annual international conference on Mobile computing and networking*. 269–280.

[93] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 210, 11 pages.

[94] Haibo Ye, Tao Gu, Xianping Tao, and Jian Lu. 2014. B-Loc: Scalable floor localization using barometer on smartphone. In *2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems*. IEEE, 127–135.

[95] Inmo Yeon and Jung-Woo Choi. 2024. RGI-Net: 3D room geometry inference from room impulse responses with hidden first-order reflections. In *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 439–443.

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

[97] Shichao Yue, Hao He, Peng Cao, Kaiwen Zha, Masayuki Koizumi, and Dina Katabi. 2022. CornerRadar: RF-based indoor localization around corners. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–24.

[98] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

[99] Yongzhao Zhang, Hao Pan, Yi-Chao Chen, Lili Qiu, Yu Lu, Guangtao Xue, Jiadi Yu, Feng Lyu, and Haonan Wang. 2023. Addressing Practical Challenges in Acoustic Sensing To Enable Fast Motion Tracking. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks (IPSN '23)*. Association for Computing Machinery, New York, NY, USA, 82–95. https://doi.org/10.1145/3583120.3586954

[100] Xiaopeng Zhao, Zhenlin An, Qingrui Pan, and Lei Yang. 2023. Nerf2: Neural radio-frequency radiance fields. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[101] Bing Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. 2017. Acoustic sensing based indoor floor plan construction using smartphones. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 519–521.

[102] Bing Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. 2017. BatMapper: Acoustic sensing based indoor floor plan construction using smartphones. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 42–55.

[103] Kaichen Zhou, Laura Dodds, Sayed Saad Afzal, and Fadel Adib. 2025. RISE: Single Static Radar-based Indoor Scene Understanding. *arXiv preprint arXiv:2511.14019* (2025).

A ACOUSTIC DATA ENGINE

A.1 Acoustic Physics

The sound propagation indoors is fundamentally determined by the interactions between acoustic waves and the geometrical structure of the space, namely the walls, floors, and large objects such as furniture. These interactions can generally be classified into a rich tapestry of acoustic effects, *i.e.*, reflection, diffusion, and absorption [41, 53]. Reflections occur as sound waves bounce off major surfaces. Diffusion further redistributes sound energy as waves interact with irregular surfaces or smaller obstacles. Collectively, these phenomena dictate the complex way in which sound reverberates and decays between any two points within the room. For a given interface, the degree of sound reflection is characterized by the r while the absorption coefficient α quantifies the fraction of incident sound energy not reflected, which is written as $\alpha = 1 - |r|^2$. These coefficients are functions of both the angle of incidence and the physical properties of the materials. According to Fresnel's law [1], the amplitude reflection coefficient for an acoustic wave incident upon a boundary at angle θ_1 relative to the normal is given by

$$r = \frac{Z \cos\theta_1 - Z_0 \cos\theta_2}{Z \cos\theta_1 + Z_0 \cos\theta_2},$$

where Z_0 and Z denote the acoustic impedance of the two adjoining media, and θ_2 is the transmission angle determined through Snell's law. In real-world scenarios, the surface is not smooth at all, therefore the sound waves will diffuse over scattered directions. To account for the scattering, the reflected energy can be further partitioned as $|r|^2 = (1 - s)|r|^2 + s|r|^2$, where s is the diffusion coefficient. Here, $(1 - s)|r|^2$ is the the portion of energy reflected specularly, while $s|r|^2$ is the portion redistributed via diffusion. The angular distribution of the diffused energy is often modeled using the Lambertian law, *i.e.*,

$$P_{\text{diff}}(\theta) = \frac{\cos\theta}{\pi}$$

where $P_{\text{diff}}(\theta)$ is the probability density function for diffuse reflection as a function of angle from the surface normal.

A.2 Existing Engine

While various simulators exist, they fundamentally fall short of the requirements for our pipeline. Traditional tools like PyRoomAcoustics [74] lack complex 3D mesh support and are entirely non-differentiable. Visual simulators like Habitat-sim [73] offer limited acoustic mesh imports and completely neglect speaker orientations. On the other hand, recent differentiable ray tracers such as Sionna [32] and its acoustic adaptation, AcoustiX [45], have emerged. However, Sionna is fundamentally designed for Electromagnetic (EM) wave propagation. To force an EM engine to simulate sound, AcoustiX

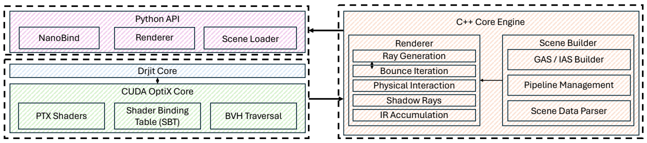


Figure 28: Framework of ARTrace

resorts to severe heuristic patches. First, to bypass EM's polarization dependence, it computes reflection coefficients from an artificially mapped permittivity, entirely discarding the angle of incidence. Second, rather than accurately tracking wave dispersion, it applies a post-hoc time-domain Gaussian or Sinc kernel to artificially sculpt the frequency envelope. Most critically, it randomly flips the phase of arriving rays, which destroys the deterministic phase coherence essential for true acoustic interference. Consequently, these compromises lead to severe macroscopic physical distortions. As shown in Fig. 10, when measuring the Energy Decay Curve (EDC) of a standard room, AcoustiX exhibits a significantly distorted decay trajectory with $r^2 = 0.90$, failing to conform to theoretical energy decay. Furthermore, across standard reverberation time (T60) measurements, AcoustiX drastically overestimates the decay time. Therefore, while designed for neural acoustic rendering, it fundamentally fails to provide the high-fidelity acoustic priors required by Loom.

A.3 Design

To address the aforementioned limitations, we introduce ARTrace. This system is a custom built differentiable ray tracing engine tailored for acoustic physics simulations. As illustrated in Fig. 28, the architecture operates on a synergy between the OptiX [67] for hardware accelerated geometric queries and the DrJit framework [33] for automatic differentiation and Just In Time compilation. The system parses scene boundary representations into a Bounding Volume Hierarchy (BVH) to accelerate spatial intersections. Concurrently, it registers material properties, specifically the frequency dependent boundary absorption and surface scattering coefficients, as explicit differentiable variables. This integration enables the computation of gradient flows from the final simulated acoustic response back to the physical material parameters of the scene during the forward rendering pass.

The simulation initiates by sampling primary ray directions from the sound source. For visible paths, the initial acoustic energy is attenuated following the inverse square law to model natural spherical geometric spreading. The arrival time is determined by the propagation distance and this continuous time value is mapped to a discrete temporal index. As the rays traverse the environment and intersect with scene geometry, the engine evaluates the material interaction at each hit point. The acoustic energy of a given ray is reduced by a fraction dictated by the associated boundary absorption coefficient. To model surface irregularities, the engine employs a probabilistic reflection model guided by the surface scattering coefficient. By drawing random floating point samples, the engine determines whether a ray undergoes a pure specular reflection, where the outgoing vector is mirrored across the surface normal, or a diffuse reflection, where a new outgoing direction is sampled within a local coordinate frame oriented around the surface normal. This mechanism simulates the complex wave scattering behavior characteristic of architectural environments.

At every boundary intersection, ARTrace evaluates the intermediate reflection contribution to all listener positions. The engine casts secondary shadow rays from the current boundary hit point towards the microphones to perform occlusion tests. If the path is clear, the physical acoustic contribution is computed by combining the residual ray energy, the attenuation from further geometric distance spreading, and the modulation from the Lambert cosine law. This cosine term is derived from the dot product between the surface normal and the normalized direction vector pointing towards the listener. The total time of flight, which accumulates the propagation time of previous bounces and the final shadow ray segment, dictates the discrete time bin for this specific multipath reflection. The resulting energy values are then populated into a global impulse response buffer using differentiable scatter and add operations. This deterministic accumulation preserves the structural envelope and the temporal phase relationship of the reflections without relying on artificial phase randomization techniques.

Simulating realistic room acoustics necessitates capturing the extensive late reverberation tail, which involves tracking thousands of successive recursive bounces. To bound the computational complexity without introducing statistical bias from arbitrary depth truncation thresholds, the engine implements a Russian Roulette path termination algorithm [70]. Once a ray trajectory exceeds a predefined minimum bounce depth limit, its survival probability is evaluated as a function of its remaining acoustic energy. Rays that fall below this probability threshold are terminated to release graphical processing unit resources. Conversely, the paths that survive this evaluation have their carried energy scaled upward by the inverse of their survival probability. This structural adjustment maintains an unbiased statistical estimator for the macroscopic thermodynamic energy decay within the simulated space. Consequently, as demonstrated in Fig. 10, the acoustic measurements synthesized by ARTrace correspond with theoretical Energy Decay Curve trajectories and standard T60 reverberation metrics.

B TRAINING IMPLEMENTATION DETAILS

Noise Scheduling Strategy: We leverage the DDIM [78] scheduler to orchestrate the diffusion process. During training, the forward process is discretized into $T = 1000$ steps. To prevent the critical geometric structures from being prematurely destroyed during the early forward steps, we adopt a cosine noise schedule `squaredcos_cap_v2` [80] with the variance parameter β progressing from 1e-4 to 1e-2. We employ Classifier-Free Guidance (CFG) [31] during the diffusion process. During training, we randomly drop the acoustic proxy conditions with a probability of 0.1. During inference, we extrapolate the predicted noise towards the conditional prediction using a scale of 10.0.

Training: The model is trained across multiple GPUs using DDP and `bf16` mixed precision. We optimize the network using AdamW, with a learning rate of 2e-4 and decay of 1e-2. We incorporate a Cosine Annealing learning rate scheduler with $\eta_{\min} = 10^{-6}$. We apply gradient norm clipping at 1.0 and maintain an Exponential Moving Average (EMA) of the model weights with a decay rate of 0.9999.

Proxy Model Loss Design: We optimize the proxy model with a multi-task objective by summing the weighted MSE loss for AoA, distance and step length estimation, reflection loss $\mathcal{L}_{\mathrm{r}}$, cross-entropy of room type classification, occupancy loss $\mathcal{L}_O$ and RIR loss $\mathcal{L}_{rir}$. Here, $\mathcal{L}_{\mathrm{r}}$ is a weighted sum of wall angle and distance MSE loss as well as the BCE loss for the wall mask; $\mathcal{L}_O$ is the sum of Dice loss, IoU loss and BCE loss for the occupancy mask; and $\mathcal{L}_{rir}$ is the sum of MSE loss of RIR waveforms as well as energy decay curves.

Geometry Prior Loss Design: We optimize the VAE using the combination of reconstruction loss, Learned Perceptual Image Patch Similarity loss [98] and KL-divergence [40] on SDF. For the reconstruction loss, it is composed of MSE loss, L1 loss, Edge-Focused MSE loss, Eikonal loss [93] and Sobel loss. We also incorporate an adversarial loss from the discriminator to further regularize the geometry prior.

C PROOF OF LEMMA 3

Proof. We compute the gradient of the score function:

$$\begin{aligned}\nabla_{z_t} s &= \nabla_{z_t} \log \frac{p_t(O_c, S_{\mathrm{obs}}|z_t)\cdot p_t(z_t)}{p_t(O_c, S_{\mathrm{obs}})}\\ &= \nabla_{z_t}\left[\log p_t(z_t) + \log p_t(O_c, S_{\mathrm{obs}}|z_t)\right]\\ &\triangleq \nabla_{z_t}\mathcal{G} + \nabla_{z_t}\mathcal{P}.\end{aligned}$$

The first term $\nabla_{z_t}\mathcal{G}$ refers to the geometry prior, which constrains the z_t to a valid floormap. Regarding the second term $\nabla_{z_t}\mathcal{P}$, although O_c is generated from S_{obs}, they represent different conditions. The proxy condition O_c is the explicit spatial prior, where the acoustic observations S_{obs} are used as implicit global acoustic features. To this end, we can further decompose the second term as $\nabla_{z_t}\mathcal{P} \approx \nabla_{z_t}\left[\log p_t(O_c|z_t) + \log p_t(S_{\mathrm{obs}}|z_t)\right] \triangleq \nabla_{z_t}\mathcal{P}_{\mathrm{spatial}} + \nabla_{z_t}\mathcal{P}_{\mathrm{match}}$. Here, the first term regularizes the spatial connection with z_t and O_c while the second term stresses the matching between S_{obs} and z_t. The optimal noise predictor $\epsilon_\theta^\star$ is known to approximate the conditional score function of the data distribution,

$$\begin{aligned}\boldsymbol{\epsilon}_\theta^*(\mathbf{z}_t, t, O_c, S_{\mathrm{obs}}) &\approx -\sigma_t \nabla_{\mathbf{z}_t} s(z_t, O_c, S_{\mathrm{obs}})\\ &\approx -\sigma_t \nabla_{\mathbf{z}_t}\left(\mathcal{G} + \mathcal{P}_{\mathrm{spatial}} + \mathcal{P}_{\mathrm{match}}\right).\end{aligned}$$

Therefore, the model jointly optimizes the three goals. □