FLOOR-PLAN-CONDITIONED GENERATIVE PRIORS FOR SPARSE SOUND-FIELD RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many spatial audio applications, such as rendering AR/VR acoustics or room correction, require knowledge of the sound field throughout a room, yet measuring it densely takes hundreds of microphone positions. With only a few microphones, many different fields fit the same measurements. Floor plans are often available and constrain the field before any measurement is taken. However, floor plans describe only the walls and doorways, not the furniture and materials that also shape the acoustic field. We therefore learn a distribution of sound fields consistent with a floor plan, using a flow-based generative prior that maps a cheap image-source field, computed from the floor plan alone, to high-fidelity wave-simulated fields. Given a few measurements, posterior sampling recovers the field of a specific room. Because the model is generative, the variance across posterior samples gives an uncertainty map along with the reconstruction. This map tracks the reconstruction error, so it can indicate where to place the next microphones. We evaluate on real multi-room apartment floor plans with wave-simulated acoustics. On floor plans unseen during training, our method reduces average reconstruction error by 5.5 dB with 128 microphones relative to kernel interpolation given the same floor-plan information, and the gain grows as microphones are added. With 16 microphones, it is at least as accurate as the same baseline with 64. Active sampling based on posterior uncertainty improves reconstruction by up to 0.7 dB over spatially uniform placement, reaching the same accuracy with 15% fewer microphones.

1 INTRODUCTION

Spatial audio aims to reproduce how sound behaves in a real space. It is used in immersive music and film, in room correction for loudspeaker systems, and in augmented and virtual reality. Doing this well requires the room's sound field: the acoustic pressure at every position and frequency, which encodes how the room's geometry and contents shape what a listener hears. Measuring this field accurately requires dense microphone measurements, often at hundreds of positions per room. Collecting them is time-consuming, and they must be repeated whenever the room changes, for example when furniture is moved.

Much of what shapes a room's sound field is known before any measurement is taken. Walls define acoustic reflections, doorways influence coupling between rooms, and the room dimensions determine its resonances. This information is recorded in the room's floor plan, which is often available from architectural drawings or real-estate listings. The floor plan is therefore free information about the sound field, and a reconstruction method that uses it should need fewer measurements.

However, the acoustic field also depends on details the floor plan does not record, such as furniture and the absorption of surface materials. Rearranging the furniture within the same floor plan produces a visibly different field (Figure 2), so a single floor plan corresponds to many valid fields. Existing methods resolve this ambiguity with measurements alone. Kernel interpolation fits a smooth field that satisfies the wave equation, and per-scene neural fields learn a representation of one room from that room's measurements. Both need dense measurements to be accurate. Neither learns what a floor plan implies about a room it has not seen, and a per-scene model must be fitted for every new room and every change of furniture.

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

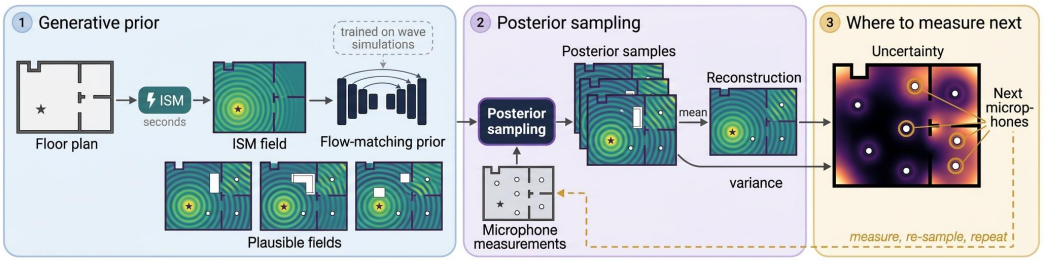


Figure 1: **Overview.** (1) An image-source (ISM) field, computed from the floor plan in seconds, conditions a flow-matching prior trained on wave-simulated fields. For a given floor plan and source (star), the prior generates plausible sound fields that differ in the furniture they imply. (2) Posterior sampling combines the prior with the microphone measurements. The posterior samples agree near the microphones and differ elsewhere, and their mean is the reconstruction. (3) Their per-location variance is the uncertainty map. A batch of new microphones goes where it is expected to reduce the uncertainty most (Section 3.3), and the loop repeats.

We take a different approach. Because one floor plan admits many fields, we treat reconstruction as a generation problem rather than a regression problem. A regression model returns a single estimate and no measure of its own uncertainty. A generative model instead represents the whole distribution of fields that a floor plan admits, so it provides both an estimate of the field and a map of where that estimate is uncertain. We train such a model on high-fidelity wave simulations, so that it learns how sound propagates in a room with acoustically relevant contents. To recover the field of a particular room, we combine this learned prior with the likelihood of the microphone measurements and sample from the resulting posterior. The prior keeps each posterior sample physically plausible, and the likelihood makes it agree with the measurements. Because the measurements enter only at inference time, one trained model serves any number and placement of microphones.

This formulation also tells us where to measure next. Where the posterior samples agree, the prior and the measurements have determined the field; where they disagree, the field is still uncertain. This uncertainty comes directly from the posterior, without a separate uncertainty model, and it depends on what the microphones recorded. The standard probabilistic counterpart of kernel interpolation, a Gaussian process, also provides a variance, but with a stationary kernel that variance depends only on where the microphones are, not on what they recorded. It therefore cannot tell which regions of a particular room are hard to predict, and placing microphones by it is no better than spreading them evenly (Section 4.3). We therefore place new microphones where the samples disagree most, update the posterior, and repeat (Figure 1). The same model both reconstructs the field and chooses the measurements.

We learn the distribution with flow matching (Lipman et al., 2023; Liu et al., 2023). The model is conditioned on a field cheaply computed from the floor plan using the image-source method (ISM). The ISM field models the walls, doorways, and direct sound, but it omits furniture, uses guessed absorption properties, and ignores wave effects such as diffraction. The model learns to map this low-fidelity field to the high-fidelity wave-simulated fields that the same floor plan produces across many furniture arrangements and materials. We evaluate on real multi-room apartment floor plans, with acoustics simulated by a wave solver, and test on floor plans unseen during training.

Contributions.

- **Generating sound fields from floor plans.** We formulate sparse sound-field reconstruction as posterior sampling under a generative prior conditioned on the floor plan, learned with flow matching from a cheap image-source field to wave-simulated fields. On unseen floor plans, our method reduces reconstruction error by 5.5 dB with 128 microphones relative to kernel interpolation given the same floor-plan information (§4.2).
- **Uncertainty fields from posterior sampling.** The variance across posterior samples tracks the reconstruction error, with a rank correlation of 0.85, and follows it far more closely than the variance of a Gaussian process fitted to the same microphones (§4.3).

- **Active sampling based on field uncertainty.** Placing new microphones where the posterior samples disagree most outperforms spatially uniform placement by up to 0.7 dB, or 15% fewer microphones, while the same rule applied to a Gaussian process's uncertainty gains nothing. The selected positions also improve kernel interpolation, which played no part in choosing them (§4.3).

2 Related Work

Sound-field reconstruction. Kernel interpolation with kernels that satisfy the Helmholtz equation is the standard method for reconstructing a sound field from sparse measurements (Ueno et al., 2018). Related approaches use Gaussian processes (Caviedes-Nozal et al., 2021), compressed sensing (Zea, 2019), and physics-informed neural networks (Olivieri et al., 2021; Karakonstantis et al., 2024); Koyama et al. (2024) review the topic. The priors in these methods are specified by hand and cannot express what a floor plan implies about the field. Gaussian-process variants provide uncertainty, but with stationary kernels the posterior variance depends only on the microphone locations, not on the signals they measure. Sundström et al. (2025) incorporate boundary geometry into a linear-Gaussian prior, whereas our prior is learned from simulated rooms.

Neural fields for room acoustics. NAF (Luo et al., 2022) and INRAS (Su et al., 2022) represent the acoustic field of a scene with a neural network fitted to that scene's measurements. AVR (Lan et al., 2024) renders impulse responses volumetrically, and INFER (Takawale et al., 2026) learns implicit frequency-response fields for confined spaces such as vehicle cabins. DiffRIR (Wang et al., 2024) reduces the number of measurements that a per-scene model needs. Other learned reconstruction methods (Lluís et al., 2020; Pezzoli et al., 2022; Xu et al., 2026) likewise return a single field for a given set of measurements. These methods produce point estimates, and the per-scene models must be fitted again whenever the room or its furniture changes.

Generative models of sound fields. Generative models have recently been applied to sound-field reconstruction, including generative adversarial networks (Fernandez-Grande et al., 2023; Karakonstantis & Fernandez-Grande, 2023), diffusion models (Miotello et al., 2024; Della Torre et al., 2025), and flow matching for field magnitudes (Erdem et al., 2026), building on score-based and flow-based generative modeling (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2025). These models are unconditional or conditioned on the source position or on measured responses. To our knowledge, none conditions the prior on room geometry. Our prior is conditioned on the floor plan, and we evaluate it on floor plans unseen during training. Liang et al. (2024) learn across rooms and return a posterior, from magnitude data and without geometric conditioning.

Posterior sampling and experimental design. We sample the posterior with diffusion posterior sampling (Chung et al., 2023), which has also been applied to audio inverse problems (Moliner et al., 2023; Lee et al., 2026). Choosing where to measure is a problem of Bayesian experimental design (Lindley, 1956; Chaloner & Verdinelli, 1995). Krause et al. (2008) derive near-optimal sensor placements for Gaussian processes and show that maximum-entropy placement concentrates sensors at the boundary of the domain, and Foster et al. (2021) amortize sequential design with a learned policy. Nolan et al. (2025) select measurements where diffusion posterior samples disagree, using a maximum-entropy criterion, for image and MRI acquisition, and Iollo et al. (2025) apply diffusion-guided Bayesian experimental design to MRI. We apply the same principle to physical sound fields, with a prior conditioned on room geometry that transfers to unseen floor plans and a bound on how many selections one set of posterior samples can support.

3 Method

We consider a room whose floor plan and source position are known. Let Ω denote the interior of the room, discretized on a grid of N nodes, and let $R_f \in \mathbb{C}^N$ denote the sound field at frequency f, the complex acoustic pressure at every node. K microphones placed at a subset of the nodes record

$$y_f \;=\; \mathcal{A}R_f + \varepsilon_f, \qquad \varepsilon_f \sim \mathcal{CN}(0, \sigma^2 I), \tag{1}$$

where $\mathcal{A} \in \{0,1\}^{K\times N}$ selects the microphone nodes. The field has far more degrees of freedom than there are microphones, so Eq. (1) has infinitely many solutions. We resolve this ambiguity with

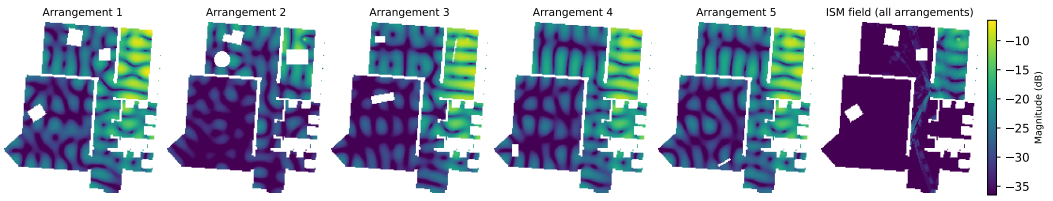


Figure 2: **One floor plan admits many sound fields.** Field magnitude at 206 Hz for five furniture arrangements of one unseen floor plan (left), and the ISM field computed from the floor plan (right). The ISM field is identical for all five arrangements, while the fields differ visibly, so the conditioning alone cannot distinguish these rooms and the measurements must. White regions are furniture and the exterior of the room.

a learned prior conditioned on the floor plan and reconstruct the field by sampling the posterior

$$p(R_f \mid y_f, b_f) \;\propto\; p(y_f \mid R_f)\, p_\theta(R_f \mid b_f), \tag{2}$$

where b_f is a field computed from the floor plan. The impulse response at every node follows from the reconstructed frequencies by an inverse discrete Fourier transform. Section 3.1 describes the prior, Section 3.2 the posterior sampler, and Section 3.3 how the posterior samples determine where to measure next.

3.1 A GENERATIVE PRIOR CONDITIONED ON THE FLOOR PLAN

The prior needs to know the geometry of the room and the position of the source. The network operates on a rectangular grid that contains the floor plan, and receives two conditioning inputs on that grid: a binary mask m equal to 1 on the nodes of Ω, and b_f, the field predicted by the image-source method (ISM) (Allen & Berkley, 1979), a standard model of room acoustics that represents reflections as contributions from mirror images of the source. We compute b_f from the floor plan and the source position with a single assumed wall absorption. The ISM field encodes the geometry, the source location, and the direct and reflected paths between them, and it does so in the same coordinates as the field to be reconstructed. It omits furniture and wave effects that image sources do not model, such as diffraction. The prior learns to map this low-fidelity field to the high-fidelity fields produced by a wave solver for the same floor plan across many furniture arrangements and absorption values. Because b_f is identical for every arrangement of a floor plan, furniture never enters the conditioning; its effect is captured by the stochasticity of the prior, and one floor plan maps to a distribution of fields (Figure 2).

We learn p_θ with rectified flow (Liu et al., 2023). A network $v_\theta(z_t, t; b_f, m)$ is trained to transport Gaussian noise z_0 to fields z_1 along the straight path $z_t = (1-t)\, z_0 + t\, z_1$ by minimizing

$$\mathcal{L}(\theta) \;=\; \mathbb{E}_{t,\, z_0,\, (z_1, b_f)} \big\| v_\theta(z_t, t; b_f, m) - (z_1 - z_0) \big\|^2, \tag{3}$$

with $t \sim \mathcal{U}[0,1]$ and $z_0 \sim \mathcal{CN}(0, I)$. In practice the network models the residual $R_f - b_f$, and b_f is added back after sampling. Each frequency satisfies its own Helmholtz equation, so we reconstruct frequencies independently with a single shared network. The network is a 2-D U-Net that receives b_f and m as additional input channels and the frequency through FiLM layers. Architecture and training details are in Appendix C.

3.2 POSTERIOR SAMPLING

We sample Eq. (2) with diffusion posterior sampling (Chung et al., 2023), adapted to rectified flow. Before sampling, a global gain and delay align the level and timing of the ISM field with the room (Appendix C.2). At each integration step, the velocity gives a one-step estimate of the clean field, $\hat{z}_1(z_t) \;=\; z_t + (1-t)\, v_\theta(z_t, t; b_f, m)$, and the sample moves along the velocity and against the gradient of the measurement misfit $\ell(z_t) = \|y_f - \mathcal{A}(b_f + \hat{z}_1(z_t))\|^2$:

$$z_{t+\Delta t} \;=\; z_t + \Delta t\, v_\theta(z_t, t; b_f, m) \;-\; \zeta\, \frac{\nabla_{z_t} \ell(z_t)}{\|\nabla_{z_t} \ell(z_t)\|}. \tag{4}$$

Repeating this for S independent initial draws gives posterior samples $R_f^{(1)}, \dots, R_f^{(S)}$, from which we form the reconstruction and a per-node uncertainty map,

$$\hat{R}_f \;=\; \frac{1}{S}\sum_{s=1}^{S} R_f^{(s)}, \qquad U(x) \;=\; \sum_f \frac{1}{S-1}\sum_{s=1}^{S} \big|R_f^{(s)}(x) - \hat{R}_f(x)\big|^2. \tag{5}$$

Because the normalized correction in Eq. (4) has magnitude ζ regardless of the misfit, a small ζ leaves the prior dominant and the measurements barely enforced, and the error at the microphone positions, where the field is known, is then no lower than elsewhere. We set ζ once on training floor plans and keep it fixed for all experiments (Appendix F.8). We use guidance in place of projection-based samplers, which enforce the data through $\mathcal{A}^\top$. For a point-sampling operator, $\mathcal{A}^\top$ returns each measurement only to the node it came from, so these samplers fit the microphones exactly and leave the field elsewhere to the unconditioned prior, while the gradient in Eq. (4) propagates through the network's Jacobian to every node (Appendix F.9).

3.3 Deciding where to measure next

The posterior samples indicate where a new microphone would be most informative. We score a candidate node ρ by the expected reduction in total posterior variance from observing it. If the posterior at frequency f is approximated as Gaussian with covariance Σ_f, observing node ρ with noise variance σ^2 reduces $\operatorname{tr}\Sigma_f$ by $\sum_x |\Sigma_{f,x\rho}|^2/(\Sigma_{f,\rho\rho}+\sigma^2)$, a rank-one update of the Gaussian covariance. Summing over frequencies gives the score

$$\Delta(\rho) \;=\; \sum_f \frac{\sum_x |\Sigma_{f,x\rho}|^2}{\Sigma_{f,\rho\rho}+\sigma^2}, \tag{6}$$

the one-step Bayesian A-optimal criterion (Chaloner & Verdinelli, 1995); the derivation is in Appendix C.5. Maximum-entropy selection ranks a node by its own variance $\Sigma_{f,\rho\rho}$ alone. The numerator of Eq. (6) also credits a node for the variance it removes at other nodes through its correlations with them. We estimate Σ_f from the deviations of the S posterior samples from their mean and evaluate Eq. (6) from these deviations directly, without forming the $N \times N$ covariance.

This estimate limits how many microphones one set of posterior samples can inform. The S deviations from the sample mean sum to zero, so the estimated covariance has rank at most $S-1$, and each noiseless observation removes one direction from its column space. After $S-1$ observations the estimated covariance vanishes and the score no longer reflects the posterior. We therefore select microphones in batches of at most $S-1$. After each batch we add the new measurements, draw new posterior samples, and recompute the score. Within a batch, we exclude candidates closer to an already selected node than a fixed fraction of the uniform microphone spacing $\sqrt{|\Omega|/K}$, so that one batch does not concentrate in a single region. The values of S, the batch size, and the exclusion fraction are given in Appendix D.

4 Evaluation

4.1 Setup

Data. We use 249 real apartment floor plans from Swiss Dwellings (Standfest et al., 2023) as a geometry source and simulate all acoustics with the differentiable wave solver j-wave (Stanziola et al., 2023). Each floor plan is rendered with five furniture arrangements and two source positions, and its wall absorption is drawn at random (Appendix B). We hold out 50 floor plans at random for testing and train on the remaining 199. No building contributes floor plans to both sets, and the unseen floor plans range from 23.6 to 77.6 m^2. The reconstruction band spans from 42 to 600 Hz.

Metric. Let $\mathcal{M}$ denote the microphone nodes. We report the normalized mean squared error over frequency bins $\mathcal{F}$ in the band and over every node without a microphone,

$$\mathrm{NMSE} \;=\; 10\log_{10}\frac{\sum_{f\in\mathcal{F}}\sum_{x\in\Omega\setminus\mathcal{M}}|\hat{R}_f(x)-R_f(x)|^2}{\sum_{f\in\mathcal{F}}\sum_{x\in\Omega\setminus\mathcal{M}}|R_f(x)|^2}, \tag{7}$$

Table 1: **Reconstruction error on 50 unseen floor plans.** Off-microphone NMSE in dB, from Eq. (7) (lower is better). All methods read the same farthest-point microphones. Matched kernel interpolation receives the same ISM field as our prior. †AVR fits one model per microphone set, so we evaluate it on 10 floor plans with one source position (Appendix E.3).

Microphones (K)	4	16	64	128	256
Textbook kernel interpolation	-0.24	-0.86	-2.46	-3.39	-4.06
Matched kernel interpolation	-1.01	-1.66	-3.23	-4.12	-4.74
INFER	$+4.84$	$+2.98$	$+2.53$	$+2.47$	$+1.95$
AVR†	$+0.10$	$+0.25$	$+0.53$	$+0.86$	$+1.83$
Ours	$\mathbf{-2.58}$	$\mathbf{-3.43}$	$\mathbf{-6.39}$	$\mathbf{-9.62}$	$\mathbf{-14.45}$
Ours − matched kernel interpolation	-1.57	-1.76	-3.16	-5.51	-9.71

Table 2: **Removing one component at a time**, with 64 microphones on the 50 unseen floor plans. Cost in dB relative to the full system (positive is worse). The last three rows use one source position per floor plan, as in the placement experiments.

Component	Replaced by	Cost (dB)
ISM field in the conditioning	free-field term, source only	$+0.31 \pm 0.10$
Generative prior	kernel interpolation, same field	$+3.16 \pm 0.28$
Guided posterior sampling	projection sampler (DDNM)	$+3.66 \pm 0.31$
Guidance weight	ζ=0.3 instead of ζ=10	$+3.69 \pm 0.32$
Uncertainty from our posterior	Gaussian-process posterior, same rule	$+0.68 \pm 0.12$
Batches within the rank bound	p=32 instead of p=6	$+0.28 \pm 0.08$
A-optimal placement	farthest-point placement	$+0.67 \pm 0.11$

so that no method is credited for reproducing its own measurements. We report paired 95% confidence intervals over the 50 unseen floor plans, rows evaluated on fewer floor plans state their own n, and we call two methods equivalent when the confidence interval of their difference lies within ± 0.5 dB. Measurements are noise-free except in the noise study of Section 4.2.

Baselines. *Textbook kernel interpolation* (Ueno et al., 2018) fits the measured pressure by kernel ridge regression with a J_0 Helmholtz kernel and is the standard reconstructor for sparse measurements. *Matched kernel interpolation* uses the same kernel but also receives the ISM field our prior is conditioned on. It interpolates the residual with respect to that field and adds the field back, so both methods receive identical information and differ only in the reconstructor. INFER (Takawale et al., 2026) and AVR (Lan et al., 2024) are per-scene neural fields, fitted to the microphones of each floor plan. We adapt INFER to complex 2-D fields and tune its capacity in its favor, because the published configuration overfits at these measurement densities (Appendix E.3). AVR fits a separate model for every microphone set, which takes about 28 minutes per floor plan and budget, so we evaluate it on 10 unseen floor plans with one source position. Unless stated otherwise, all methods read the same microphones, placed by farthest-point sampling, which spreads them evenly over the floor plan. Hyperparameters are listed in Appendix D and the full protocol in Appendix E.

4.2 Reconstruction

Table 1 reports reconstruction error across microphone budgets. Our method outperforms matched kernel interpolation on every unseen floor plan at every budget, by 5.51 ± 0.46 dB with 128 microphones. The margin grows with the budget, from 1.57 dB with 4 microphones to 5.51 dB with 128 and 9.71 dB with 256. A learned prior usually helps most when measurements are scarce and matters less as they accumulate. Here the opposite holds, which suggests that the floor plan and the measurements carry complementary information. Each added microphone is also worth more to our method: from 64 to 128 microphones, its error falls by 3.2 dB, against 0.9 dB for matched kernel interpolation. Table 2 removes one component at a time; with 64 microphones, every removal increases the error. Across three training seeds, the margin with 128 microphones is 5.51, 5.74 and 5.97 dB; the reported model is the least accurate of the three (Appendix F.3).

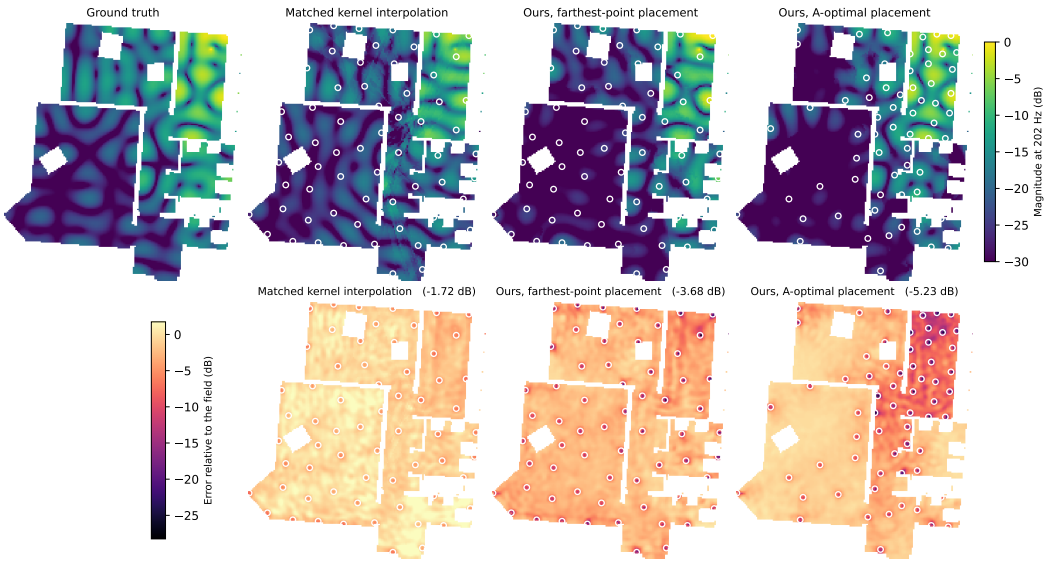


Figure 3: **Reconstruction on an unseen floor plan** with 64 microphones (circles). Top: field magnitude at 202 Hz for the ground truth, matched kernel interpolation, and our posterior mean with farthest-point and A-optimal placement. Bottom: broadband error at each node, with the off-microphone NMSE in each panel title. White regions are furniture.

The gain translates into fewer microphones (Appendix F.1). With 16 microphones, our method is at least as accurate as matched kernel interpolation with 64 (0.20 ± 0.10 dB better), and with 64 it is 2.27 ± 0.28 dB more accurate than matched kernel interpolation with 128. It matches textbook kernel interpolation, which does not use the floor plan, with eight times fewer microphones: at 16 against 128, the paired difference lies between -0.18 and $+0.10$ dB, within this equivalence margin. These results hold under measurement noise. At 20 dB SNR, the margin over matched kernel interpolation narrows by 0.06 ± 0.02 dB with 64 microphones; at 10 dB, it narrows by 0.50 ± 0.10 dB with 64 microphones and 1.34 ± 0.23 dB with 128. In both cases our method still outperforms matched kernel interpolation on every floor plan at every budget from 4 to 128 microphones (Appendix F.2).

Per-scene neural fields are designed for a different regime. They learn one room from its own measurements and are typically fitted to densely sampled scenes; INFER, for example, was developed for vehicle cabins. At the budgets studied here, the microphones are sparse relative to the wavelength: with 128 microphones in a 50 m^2 apartment, neighboring microphones are about 0.6 m apart, farther than half a wavelength at frequencies above $\sim$270 Hz, so the measurements alone do not determine the field. Our prior supplies the missing information from other rooms. A per-scene model has no such prior and must infer the field between microphones from those microphones alone. INFER reproduces the measured positions closely, with errors of -66 to -15 dB at the microphones, but generalizes poorly away from them (Table 1). AVR improves on INFER by 2.32 ± 0.55 dB on the 10 floor plans where we evaluate it with 64 microphones, and our method is a further 6.69 ± 1.64 dB more accurate there. Figure 3 shows reconstructions on one unseen floor plan.

4.3 Active sampling

Posterior variance as an error estimate. Choosing where to measure requires an uncertainty estimate that reflects the specific room. On the 50 unseen floor plans with 64 microphones, the rank correlation between the posterior variance $U(x)$ of Eq. (5) and the true off-microphone error is 0.85 ± 0.02. A Gaussian process fitted to the same microphones with the same ISM field reaches 0.13 ± 0.04 when scored the same way, and 0.14 ± 0.04 without the ISM field. It uses the J_0 kernel of each frequency, with its regularization fitted per frequency by marginal likelihood. With a stationary kernel, that variance is set by the microphone positions alone: fitting the regularization changes its scale, not where it is large. Our posterior variance instead depends on the recorded values and on the learned prior (Figure 4).

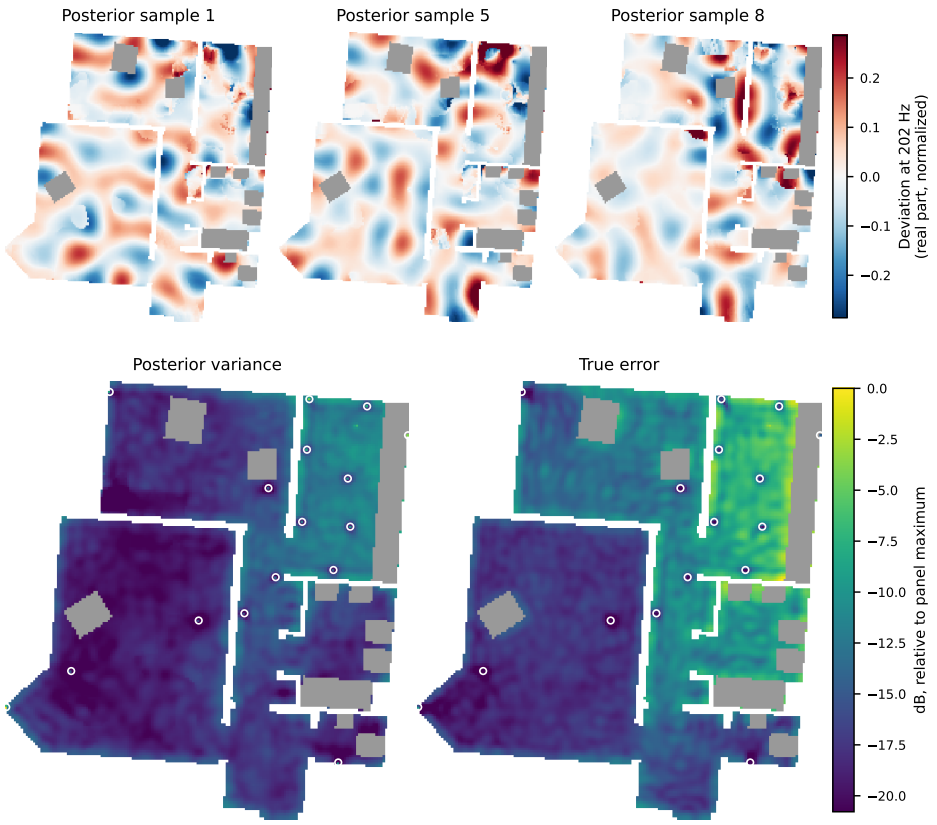


Figure 4: **Posterior samples and uncertainty** on an unseen floor plan with 16 microphones selected by A-optimal placement (circles). Top: three posterior samples, shown as their deviation from the posterior mean. Bottom: the posterior variance $U(x)$ beside the true error; their rank correlation here is 0.74. Gray regions are furniture.

A-optimal placement. Selecting microphones by Eq. (6) in batches of p=6 outperforms farthest-point placement by 0.67 ± 0.11 dB with 64 microphones, in a comparison whose test statistic and success criterion we fixed before running it. The gain grows with the budget, from 0.26 ± 0.04 dB with 10 microphones, and exceeds 0.5 dB on average from 28 microphones onward (Figure 5b). Equivalently, A-optimal placement reaches the accuracy of farthest-point placement with 15% fewer microphones: farthest-point placement needs 75 ± 2 microphones to match it with 64. With the same rule, batches and warm start, the covariance of the Gaussian process above gives no gain over farthest-point placement (0.00 ± 0.08 dB), whereas maximum entropy on our posterior is equivalent to Eq. (6) (difference 0.05 ± 0.08 dB). The gain therefore comes from our posterior's uncertainty; the choice of rule matters little. The two rules spend their budgets differently. Farthest-point placement puts $40.1 \pm 1.7\%$ of its microphones in the half of each floor plan where the ISM field is already most accurate, and A-optimal placement puts $24.9 \pm 1.7\%$ there, concentrating the rest where the floor plan alone predicts the field worst (Figure 5a). Exceeding the rank bound of Section 3.3 gives up part of this gain: selecting p=32 microphones per set of posterior samples instead of p=6 costs 0.28 ± 0.08 dB with 64 microphones, about 40% of the improvement over farthest-point placement (Table 2).

Cross-reconstructor evaluation. Because the same posterior selects the microphones and reconstructs from them, A-optimal placement could appear to win by agreeing with its own reconstructor. We therefore reconstruct from the same selected positions with matched kernel interpolation, which played no part in choosing them. Kernel interpolation retains about half of the gain in decibels, 0.34 ± 0.05 dB with 64 microphones; with farthest-point placement it would need about 81 microphones to reach the same accuracy. The selected positions are therefore more informative in their

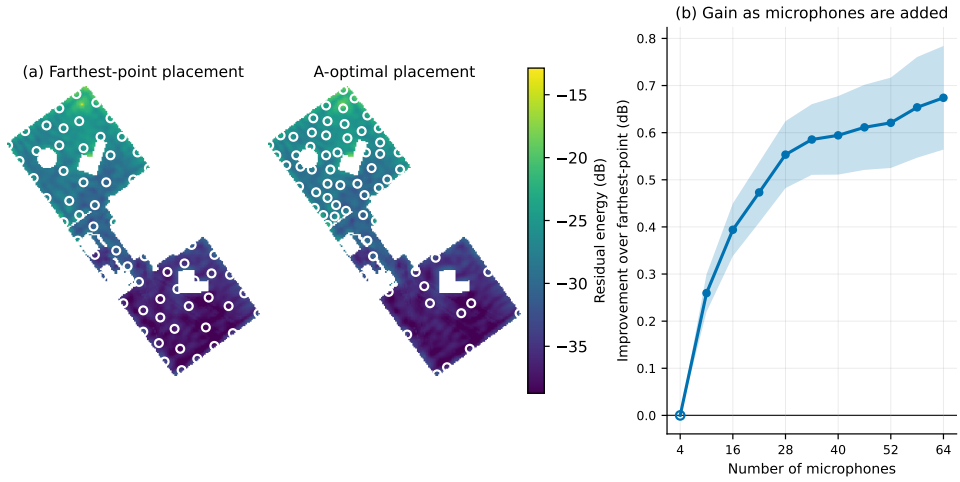


Figure 5: **A-optimal placement.** (a) Microphone positions from farthest-point and A-optimal placement on an unseen floor plan, over the residual energy relative to the ISM field (bright where the floor plan alone predicts the field worst). Circles mark 64 microphones selected in batches of $p{=}6$; white regions are furniture. (b) Improvement of A-optimal over farthest-point placement as microphones are added (mean and 95% CI, 50 unseen floor plans); both rules share the first four microphones.

own right, and the other half of the gain comes from how our model uses them. With 16 microphones, most of the gain comes from the positions (Appendix F.6).

5 Discussion and Limitations

The approach requires only an approximate forward model that is cheap to evaluate, so it should carry over to other field-reconstruction problems, such as radio-frequency coverage mapping. Our evaluation uses real floor plans with acoustics simulated by a wave solver as ground truth. The solver agrees with analytic free-field solutions and satisfies reciprocity closely (Appendix B.2), which bounds its numerical error; validating the approach on measured rooms is the most important next step. The study is two-dimensional, and the prior relies on an accurate floor plan: offsetting every wall outward by 5 cm costs 1.70 ± 0.18 dB with 64 microphones, and at 20 cm matched kernel interpolation becomes more accurate; moving each wall independently by a random offset with a 5 cm standard deviation costs 0.81 ± 0.18 dB (Appendix F.4). The acquisition criterion is greedy and scores one step ahead under a Gaussian approximation of the posterior; non-myopic design (Foster et al., 2021) could plan several measurements jointly. Extending the method to three-dimensional rooms and higher frequencies requires larger simulations and networks, while the formulation remains the same.

6 Conclusion

We reconstruct sound fields by sampling from a learned distribution conditioned on a floor plan. A flow-based prior maps an inexpensive image-source field, computed from the floor plan, to high-fidelity, wave-simulated fields. Posterior sampling combines this prior with sparse microphone measurements to recover the actual field. The variance of the posterior samples tracks the reconstruction error, and we use it to decide where to place the next microphones. On real apartment floor plans unseen during training, the method reduces reconstruction error by 5.5 dB with 128 microphones relative to kernel interpolation given the same floor-plan information, and with 16 microphones it is at least as accurate as that baseline with 64. Uncertainty-driven placement reaches the accuracy of spatially uniform placement with 15% fewer microphones. Treating reconstruction as a generation task yields both the field estimate and the uncertainty that guides the next measurements.

Reproducibility statement

All results come from simulation with a differentiable wave solver, with fixed seeds. Appendix C describes the method in full, Appendix D lists every hyperparameter, and Appendix E specifies the evaluation protocol and each baseline. Every number in the main text is recomputed from stored evaluation outputs by a verification script.

Use of large language models

We used large language models and generative AI tools to assist with code development, experiment orchestration, and editing of the manuscript. Figure 1 was generated with an image-generation model from a layout specified by the authors. We did not use generative AI tools to create data: all training and evaluation data come from our wave simulations. The authors designed the method and experiments, verified all results and claims, and take full responsibility for the content.

References

Michael Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

Diego Caviedes-Nozal, Nicolai A. B. Riis, Franz M. Heuchel, Jonas Brunskog, Peter Gerstoft, and Efren Fernandez-Grande. Gaussian processes for sound field reconstruction. *The Journal of the Acoustical Society of America*, 149(2):1107–1119, 2021. doi: 10.1121/10.0003497.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. doi: 10.1214/ss/1177009939.

Hyungjin Chung, Jeongsol Kim, Michael T McCann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023.

Sagi Della Torre, Mirco Pezzoli, Fabio Antonacci, and Sharon Gannot. DiffusionRIR: Room impulse response interpolation using diffusion models. *arXiv preprint arXiv:2504.20625*, 2025.

Ege Erdem, Shoichi Koyama, Tomohiko Nakamura, Orchisama Das, and Zoran Cvetković. SF-Flow: Sound field magnitude estimation via flow matching guided by sparse measurements. *arXiv preprint arXiv:2605.10398*, 2026.

Efren Fernandez-Grande, Xenofon Karakonstantis, Diego Caviedes-Nozal, and Peter Gerstoft. Generative models for sound field reconstruction. *The Journal of the Acoustical Society of America*, 153(2):1179–1190, 2023. doi: 10.1121/10.0016896.

Adam Foster, Desi R Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *ICML*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Jacopo Iollo, Geoffroy Oudoumanessah, Carole Lartizien, Michel Dojat, and Florence Forbes. Active MRI acquisition with diffusion guided Bayesian experimental design. *arXiv preprint arXiv:2506.16237*, 2025.

Xenofon Karakonstantis and Efren Fernandez-Grande. Generative adversarial networks with physical sound field priors. *The Journal of the Acoustical Society of America*, 154(2):1226–1238, 2023. doi: 10.1121/10.0020665.

Xenofon Karakonstantis, Diego Caviedes-Nozal, Antoine Richard, and Efren Fernandez-Grande. Room impulse response reconstruction with physics-informed deep learning. *The Journal of the Acoustical Society of America*, 155(2):1048–1059, 2024. doi: 10.1121/10.0024750.

Shoichi Koyama, Juliano G. C. Ribeiro, Tomohiko Nakamura, Natsuki Ueno, and Mirco Pezzoli. Physics-informed machine learning for sound field estimation: Fundamentals, state of the art, and challenges. *IEEE Signal Processing Magazine*, 41(6):60–71, 2024. doi: 10.1109/MSP.2024.3465896.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008. URL https://jmlr.org/papers/v9/krause08a.html.

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

Kyung Yun Lee, Nils Meyer-Kahlen, Vesa Välimäki, and Sebastian J. Schlecht. Solving room impulse response inverse problems using flow matching with analytic Wiener denoiser. *arXiv preprint arXiv:2602.00652*, 2026.

Zining Liang, Wen Zhang, and Thushara D. Abhayapala. Sound field reconstruction using neural processes with dynamic kernels. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):13, 2024. doi: 10.1186/s13636-024-00333-x.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Francesc Lluís, Pablo Martínez-Nuevo, Martin Bo Møller, and Sven Ewan Shepstone. Sound field reconstruction in rooms: Inpainting meets super-resolution. *The Journal of the Acoustical Society of America*, 148(2):649–659, 2020. doi: 10.1121/10.0001687.

Andrew Luo, Yilun Du, Michael J Tarr, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. In *NeurIPS*, 2022.

Federico Miotello, Luca Comanducci, Mirco Pezzoli, Alberto Bernardini, Fabio Antonacci, and Augusto Sarti. Reconstruction of sound field through diffusion models. In *Proc. IEEE ICASSP*, pp. 1476–1480, 2024. doi: 10.1109/ICASSP48485.2024.10446761.

Eloi Moliner, Jaakko Lehtinen, and Vesa Välimäki. Solving audio inverse problems with a diffusion model. In *Proc. IEEE ICASSP*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095637.

Oisin Nolan, Tristan S. W. Stevens, Wessel L. van Nierop, and Ruud J. G. van Sloun. Active diffusion subsampling. *Transactions on Machine Learning Research*, 2025. arXiv:2406.14388.

Marco Olivieri, Mirco Pezzoli, Fabio Antonacci, and Augusto Sarti. A physics-informed neural network approach for nearfield acoustic holography. *Sensors*, 21(23):7834, 2021. doi: 10.3390/s21237834.

Mirco Pezzoli, Davide Perini, Alberto Bernardini, Federico Borra, Fabio Antonacci, and Augusto Sarti. Deep prior approach for room impulse response reconstruction. *Sensors*, 22(7):2710, 2022. doi: 10.3390/s22072710.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

Matthias Standfest, Michael Franzen, Yvonne Schröder, Luis Gonzalez Medina, Yarilo Villanueva Hernandez, Jan Hendrik Buck, Yen-Ling Tan, Milena Niedzwiecka, and Rachele Colmegna. Swiss Dwellings: A large dataset of apartment models including aggregated geolocation-based simulation results covering viewshed, natural light, traffic noise, centrality and geometric analysis, 2023. Zenodo, version 3.0.0.

Antonio Stanziola, Simon R Arridge, Ben T Cox, and Bradley E Treeby. j-Wave: An open-source differentiable wave simulator. *SoftwareX*, 22:101338, 2023.

Kun Su, Mingfei Chen, and Eli Shlizerman. INRAS: Implicit neural representation for audio scenes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8144–8158, 2022.

David Sundström, Filip Elvander, and Andreas Jakobsson. Boundary-informed sound field reconstruction. In *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 81–85, 2025.

Harshvardhan C Takawale, Nirupam Roy, and C Phillip Brown. INFER: Learning implicit neural frequency response fields for confined acoustic environments. In *Forty-third International Conference on Machine Learning*, 2026.

Natsuki Ueno, Shoichi Koyama, and Hiroshi Saruwatari. Kernel ridge regression with constraint of Helmholtz equation for sound field interpolation. In *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 436–440, Tokyo, Japan, September 2018.

Mason Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11790–11799, 2024.

Shaoheng Xu, Chunyi Sun, Jihui Zhang, Prasanga N. Samarasinghe, and Thushara D. Abhayapala. RIR-Former: Coordinate-guided transformer for continuous reconstruction of room impulse responses. In *Proc. IEEE ICASSP*, pp. 15312–15316, 2026.

Elias Zea. Compressed sensing of impulse responses in rooms of unknown properties and contents. *Journal of Sound and Vibration*, 459:114871, 2019. doi: 10.1016/j.jsv.2019.114871.

A Overview

Additional visualizations, code and the dataset are available at `https://anon61823-star.github.io`. Appendix B describes the apartment corpus and validates the wave solver. Appendix C specifies the method: the ISM field, the residual target, the measurement gauge, the network, the sampler, and the derivations behind Eq. (6) and the bound on batch size. Appendix D collects every hyperparameter in one table. Appendix E gives the evaluation protocol and the configuration of each baseline. Appendix F reports the full results behind the main-text numbers. It also includes a control, which the main text does not cite, showing that the support of the ISM field at evaluation does not affect accuracy (Appendix F.10).

B Corpus and simulation

This appendix describes the apartment corpus behind every result and the checks on the wave solver.

B.1 The apartment corpus

We take 249 apartment floor plans from Swiss Dwellings (Standfest et al., 2023), which provides geometry only and contains no acoustic measurements. All acoustics are simulated with the differentiable wave solver j-wave (Stanziola et al., 2023). The acoustic interior of each apartment is the union of its indoor areas, with doors left open, windows closed, and the fixed fixtures (bath, shower, sink, toilet, kitchen and built-in cabinets) kept at their recorded positions. We keep a floor plan only if its rooms, joined through their doors, form one connected interior with no enclosed column and no passage narrower than 0.3 m; its area lies between 20 and 80 m^2; its outline does not duplicate another floor plan; and in every furniture arrangement the fixtures and furniture leave at least 99% of the interior connected. Among the floor plans that pass, we sample evenly across the area range. Each floor plan is rendered with five furniture arrangements and two source positions, and twice more without movable furniture, for 2,988 renders in total. The fixed fixtures are the same in every arrangement of a floor plan. Each arrangement adds one to three movable items, placed by seeded rejection sampling under wall-clearance, non-overlap and 20% floor-occupancy constraints, with a check that no item disconnects the interior. Wall absorption is drawn once per floor plan from

$\mathcal{U}[0.35, 0.50]$. Renders use a sampling rate of 2400 Hz and $N{=}512$ samples on a lattice with spacing 0.0572 m. We model the 128 real-FFT bins from 4.69 to 600 Hz. The whitening filter is zero below 40 Hz, so the lowest eight bins carry no energy and the effective band is 42.2 to 600 Hz. We hold out 50 floor plans at random and train on the other 199. No building contributes floor plans to both sets, and the unseen floor plans range from 23.6 to 77.6 m^2.

B.2 Solver validation

We check the wave solver against analytic solutions at three probe frequencies. Exchanging the source and a receiver reproduces the field to -99.31 dB, consistent with acoustic reciprocity. Against the two-dimensional free-field Green's function, the solver's error is -79.09, -75.77 and -72.60 dB at 250, 500 and 1000 Hz, and the fitted exponent of amplitude decay with distance matches the theoretical $-1/2$ to within 3×10^{-3} at all three. Two probes lie inside the modeled band and the 1000 Hz probe lies above it.

C Method details

This appendix specifies what Section 3 summarizes: how the ISM field is computed, the residual target, the measurement gauge, the network and its training, the sampler, and the derivations behind Eq. (6) and the bound on batch size. Table 3 lists every numerical setting.

C.1 The ISM field

The ISM field b_f is computed from the floor plan alone. We apply the image-source method to the floor-plan polygon up to reflection order three, keep only images whose specular path to the receiver is unobstructed, so that non-convex floor plans are handled, and sum the two-dimensional free-field Green's function, a zeroth-order Hankel function, over the source and its images at each frequency bin. The true wall absorption is unknown at deployment, so every wall uses a guessed absorption $\hat{\alpha}{=}0.5$, while the simulated absorption varies per floor plan over $[0.35, 0.50]$. The field is computed once per floor plan on the empty interior and passed through the same whitening filter as the simulated fields.

C.2 Residual target and measurement gauge

The prior models the residual $r_f = R_f - b_f$ between the simulated field R_f and the gauged ISM field, which is the part of the field that the floor plan does not predict: the effect of furniture, the error in the guessed absorption, and the error of the image-source approximation. The reconstruction is $\hat{R}_f = b_f + \hat{r}_f$, where $\hat{r}_f$ is the mean of the posterior residual samples, and every NMSE compares $\hat{R}_f$ with R_f.

Before sampling, a gain and a delay align the ISM field with the room: b_f is replaced by $a\, e^{-2\pi i f \tau} b_f$. In the reported experiments, the gain and delay were fitted to the simulated empty-room field at every node, in training and at evaluation. The delay τ maximizes the cross-correlation between the ISM field and that field, summed over nodes and frequency bins and refined to a fraction of a sample, and the gain a is the least-squares gain at that delay. Fitting them to the K microphones instead, which uses no information unavailable at deployment, changes our NMSE by $+0.06 \pm 0.04$ dB with 16 microphones and $+0.00 \pm 0.01$ dB with 128, and matched kernel interpolation by at most 0.04 dB from 16 microphones upward. With 4 microphones the fitted gain is less reliable, and our NMSE changes by $+0.42 \pm 0.43$ dB.

C.3 Network and training

The prior is a U-Net with base width 96 and 2.70M parameters on a 305×320 grid. Its five input channels are the current state of the flow (real and imaginary parts), the interior mask, and the gauged ISM field (real and imaginary parts), and it outputs the two-channel velocity. The frequency enters as a 64-dimensional sinusoidal code of $f/600$ Hz applied through FiLM layers, and the flow time through the network's own time embedding. Residuals and ISM fields are each divided by one global scale. We train with the rectified-flow objective, starting from standard Gaussian noise. Each

training example is one render, meaning one furniture arrangement and one source position, and one frequency bin, both drawn uniformly and independently.

C.4 SAMPLER

We integrate the flow from $t{=}0$ to $t{=}1$ in $T{=}20$ Euler steps from standard Gaussian noise and draw $S{=}8$ posterior samples per reconstruction. At each step the state moves along the velocity and against the gradient of the measurement misfit (Eq. (4)). The gradient is normalized to unit norm, so the guidance weight $\zeta{=}10$ sets the size of the correction independently of the scale of the misfit. The reconstruction is the mean of the S posterior samples, and their spread gives the variance map used for acquisition.

C.5 DERIVATION OF THE ACQUISITION SCORE AND THE BATCH BOUND

This subsection derives Eq. (6) and the bound of Section 3.3 on how many microphones one set of posterior samples can select.

Fix a frequency bin and let $r \in \mathbb{C}^P$ be the residual on the P candidate nodes. Given posterior samples $r^{(1)}, \dots, r^{(S)}$ with mean $\bar{r}$, write $d^{(s)} = r^{(s)} - \bar{r}$ and $\Sigma = \frac{1}{S}\sum_s d^{(s)} d^{(s)*}$ for the sample covariance, whose diagonal is the variance map. Observing node ρ returns $y = r(\rho) + \varepsilon$ with noise variance σ^2. The simulated measurements are noise-free, so σ^2 acts as a ridge, fixed at 20 dB below the root-mean-square deviation of the samples.

Proposition 1 (One-step Bayesian A-optimality)**.** *If the posterior is Gaussian with covariance Σ and node ρ is observed with noise variance $\sigma^2 > 0$, the expected reduction in total posterior variance* $\operatorname{tr}\Sigma$ *is*

$$\Delta(\rho) \;=\; \frac{\sum_x |\Sigma_{x\rho}|^2}{\Sigma_{\rho\rho} + \sigma^2}, \tag{8}$$

so $\arg\max_\rho \Delta(\rho)$ *is the greedy one-step Bayesian A-optimal design, which minimizes the expected posterior squared error (Chaloner & Verdinelli, 1995). It differs from maximizing expected information gain, which for one linear-Gaussian observation equals* $\log(1 + \Sigma_{\rho\rho}/\sigma^2)$ *for complex Gaussian fields and so ranks nodes by their own variance, as maximum entropy does.*

Proof. For a linear-Gaussian observation the posterior covariance is the rank-one downdate $\Sigma' = \Sigma - \Sigma e_\rho (\Sigma_{\rho\rho} + \sigma^2)^{-1} e_\rho^* \Sigma$. Taking traces and using $\Sigma_{\rho x} = \Sigma_{x\rho}^*$ gives $\operatorname{tr}\Sigma - \operatorname{tr}\Sigma' = \sum_x |\Sigma_{x\rho}|^2/(\Sigma_{\rho\rho} + \sigma^2)$. □

Treating frequency bins as independent, the reduction over the band is the sum of Eq. (8) over bins, which is Eq. (6). The implementation computes each bin's term from the deviations $d^{(s)}$ without forming Σ, and sums over all bins. Scaling Σ by a constant does not change the $\arg\max$, so the choice between $1/S$ and $1/(S-1)$ in the sample covariance does not affect the selected node.

Proposition 2 (One set of posterior samples supports at most $S-1$ selections)**.** *The deviations satisfy $\sum_s d^{(s)} = 0$, so* $\operatorname{rank}\Sigma \le S - 1$. *Each noiseless observation removes one direction from the column space of Σ. After $S-1$ observations $\Sigma = 0$, every $\Delta(\rho)$ vanishes, and the* $\arg\max$ *is decided by what remains below the ridge.*

The bound is $S-1$ and not S because the sample mean is subtracted. Running the implemented selector on sets of $S = 4$, 8 and 16 samples, the remaining variance collapses after exactly $S-1$ selections each time. With $S{=}8$, it falls from 0.13 of its initial value after six selections to 6.5×10^{-5} after seven. The bound is why we select microphones in batches of $p{=}6 \le S-1$ and redraw the posterior samples after each batch. Selecting $p{=}32$ microphones from one set instead costs 0.28 ± 0.08 dB at $K{=}64$ on the apartment corpus (Table 2).

Remark 1 (Scope). The Gaussian surrogate and the rank-deficient covariance estimated from $S{=}8$ samples are approximations; the true posterior is neither Gaussian nor rank-deficient. Both propositions are also one-step results. The $(1 - 1/e)$ guarantee of Krause et al. (2008) for greedy selection relies on the submodularity of mutual information, and the variance-reduction objective of Eq. (8) is not submodular in general, so no such guarantee holds for the batch rule.

D HYPERPARAMETERS

Table 3 lists every setting needed to rerun the system on the apartment corpus. All reported results use these values unless a result states otherwise; the guidance weight, the number of posterior samples, the number of integration steps and the batch size are varied in Appendix F.

Table 3: **Hyperparameters.** Values used for every result on the apartment corpus.

Group	Setting	Value
Data	Sampling rate, length	2400 Hz, N=512 samples
	Modeled bins	128 real-FFT bins, 4.69 to 600 Hz
	Effective band	42.2 to 600 Hz
	Lattice spacing	0.0572 m
	Training set	199 floor plans, 1,990 renders
ISM field	Reflection order	3
	Wall absorption (guessed)	$\hat{\alpha} = 0.5$
	Gauge	gain and delay, fitted to the simulated empty-room field (Appendix C.2)
Network	Architecture	U-Net on a 305×320 grid
	Base width, parameters	96, 2.70M
	Input, output channels	5 (state, mask, ISM field), 2 (velocity)
	Frequency conditioning	64-dim. sinusoidal code of $f/600$ Hz, FiLM
Training	Objective	rectified flow on the residual
	Optimizer	AdamW, weight decay 0, no gradient clipping
	Learning rate	2×10^{-3}, 1,000-step linear warmup, cosine decay
	Batch size, steps	16, 40,000
	EMA decay	0.999 (EMA weights used at inference)
Sampling	Integration	T=20 Euler steps from standard Gaussian noise
	Posterior samples	S=8
	Guidance weight	ζ=10, unit-normalized gradient
Acquisition	Score	Eq. (6), summed over all 128 bins
	Ridge σ^2	20 dB below the RMS deviation of the samples
	Batch size	p=6 ($\leq S - 1$)
	Exclusion radius	$\frac{1}{\sqrt{\pi}}\sqrt{\lvert\Omega\rvert/K}$
	Candidates	all interior nodes without a microphone
	Initial microphones	first 4 by farthest-point placement

E EVALUATION PROTOCOL AND BASELINES

This appendix specifies how every number in Section 4 is computed and how each baseline in Table 1 is configured.

E.1 METRIC AND STATISTICS

NMSE follows Eq. (7), with the sum over f running over the 128 modeled bins. We compute it on the time-domain response assembled by an inverse real FFT over those bins. The modeled bins exclude the zero-frequency and Nyquist bins, so Parseval's relation weights every bin equally and the time-domain value equals the frequency-domain sum exactly. The microphone nodes are excluded from both sums.

The unit of analysis is the floor plan. We average a floor plan's two source positions before pairing, and pair two methods only over floor plans evaluated with the same source positions in both. Each reported difference is a paired mean with a 95% confidence interval from the t distribution with $n-1$ degrees of freedom, where n is the number of floor plans (50). We also report on how many floor plans the difference favors each method. Two methods are called equivalent when the confidence interval of their difference lies within ± 0.5 dB.

E.2 MICROPHONE PLACEMENT AND THE KERNEL BASELINES

Unless a result states otherwise, the K microphones are placed by farthest-point sampling on the interior lattice, and every method reads the same microphone set; we check that the sets are identical before comparing any two methods.

Both kernel ridge regression baselines interpolate each frequency bin separately with the J_0 Helmholtz kernel $J_0(k\|x - x'\|)$, $k = 2\pi f/c$ with $c = 343$ m/s, and set the ridge by the same rule, $\lambda = \max\big(10^{-2},\, 0.2\,\lambda_{\max}(K_{\mathcal{M}\mathcal{M}})\big)$, where $K_{\mathcal{M}\mathcal{M}}$ is the kernel matrix among the microphones. Textbook kernel interpolation (Ueno et al., 2018) interpolates the measured pressure. Matched kernel interpolation subtracts the gauged ISM field at the microphones, interpolates the residual, and adds the field back, so it receives exactly the information our prior receives. It outperforms textbook interpolation at every budget (Table 1), so the main text compares against it.

A fixed ridge rule does not handicap the kernel method. A Gaussian process with the same kernel, whose noise-to-signal ratio is inferred per floor plan and per frequency bin by maximizing the marginal likelihood (Caviedes-Nozal et al., 2021), is less accurate than textbook interpolation on the same microphones by 1.03 ± 0.07 dB at $K{=}64$ (on 50 of 50 floor plans) and 0.23 ± 0.08 dB at $K{=}128$. Compressed sensing over an over-complete plane-wave dictionary, with its sparsity weight chosen on a training floor plan, is less accurate by 0.65 ± 0.05 dB at $K{=}64$ (50 of 50) and 0.12 ± 0.04 dB at $K{=}128$.

E.3 PER-SCENE NEURAL FIELDS

INFER (Takawale et al., 2026) fits an implicit frequency-response field to the measurements of one scene. We port it to two-dimensional complex fields and fit it per floor plan to the same microphones, in total-field mode without the ISM field, as it was published. Its capacity was chosen by a sweep on a training floor plan: we use 0.34M parameters (eight hash-grid levels with 2^{12} entries each, width 256, depth 6) trained for 3,000 iterations at learning rate 5×10^{-3}, because the published configuration of about 3.5M parameters overfits at these measurement densities. We omit three components that do not apply here: the Kramers–Kronig consistency constraint, which assumes a causal full-band response while our fields are whitened and band-limited; the perceptual spectral weighting, which is defined over 20 Hz to 20 kHz; and orientation conditioning, since our sources and receivers are omnidirectional. INFER's error at the microphones rises from -66.5 dB at $K{=}4$ to -15.1 dB at $K{=}128$.

AVR (Lan et al., 2024) is a volumetric acoustic renderer that fits one model per microphone set. We run it with the speed of sound (343 m/s) and sampling rate (2400 Hz) of our renders, a 512-tap output, 80 samples per ray, three hidden layers in each of its networks, and 40,000 iterations at learning rate 5×10^{-4}. Its loss weights time-domain error by 50, amplitude by 4, the spectrum by 2, and angle, energy and multi-resolution STFT errors by 1. One fit takes 1,692 s, about 28 minutes, and a fit serves only one microphone set, so we evaluate AVR on 10 unseen floor plans with one source position.

E.4 GAUSSIAN-PROCESS UNCERTAINTY

The uncertainty comparison of Section 4.3 fits a Gaussian process with the same J_0 kernel to the residual with respect to the same gauged ISM field, on the same microphones. Its only free hyperparameter, the noise-to-signal ratio, is chosen per frequency bin by maximizing the marginal likelihood over a grid of 41 values from 10^{-6} to 10^{2}; the kernel's length scale is set by $k = 2\pi f/c$. We sum the posterior variance over bins, sum the squared off-microphone error over bins, and report the Spearman rank correlation between the two over the off-microphone nodes, the same expression we apply to our posterior samples. The Gaussian process reaches 0.13 ± 0.04 at $K{=}64$ and 0.05 ± 0.04 at $K{=}16$ over the 50 unseen floor plans. Our posterior reaches 0.85 ± 0.02 and 0.81 ± 0.02 at the same budgets.

F ADDITIONAL RESULTS

This appendix gives the full data behind the results of Section 4. Every result is on the 50 unseen floor plans of the apartment corpus unless it states otherwise, and every difference is a paired mean with a 95% confidence interval (Appendix E.1). The last subsection is a control that the main text does not cite.

F.1 MICROPHONE SAVINGS

Table 4 lists every cross-budget comparison behind the savings reported in Section 4.2.

Table 4: **Cross-budget comparisons.** Paired difference in off-microphone NMSE, ours minus the baseline (negative favors ours), with the number of floor plans on which ours is more accurate; n=50.

Comparison	Microphone ratio	Difference (dB)	Floor plans won
Ours at 4 vs. matched at 16	4×	-0.92 ± 0.08	50/50
Ours at 16 vs. matched at 64	4×	-0.20 ± 0.10	38/50
Ours at 64 vs. matched at 128	2×	-2.27 ± 0.28	50/50
Ours at 16 vs. matched at 128	8×	$+0.69 \pm 0.10$	1/50
Ours at 4 vs. textbook at 16	4×	-1.72 ± 0.14	50/50
Ours at 16 vs. textbook at 64	4×	-0.97 ± 0.14	49/50
Ours at 64 vs. textbook at 128	2×	-3.00 ± 0.30	50/50
Ours at 16 vs. textbook at 128	8×	-0.04 ± 0.14	28/50

The fourth row shows where the savings end. With eight times fewer microphones our method is less accurate than matched kernel interpolation, by 0.69 ± 0.10 dB. Against textbook interpolation, which does not use the floor plan, the same ratio gives a difference whose interval lies inside the ± 0.5 dB equivalence margin.

F.2 MEASUREMENT NOISE

We add complex Gaussian noise to the measurements. Its power is set relative to the root-mean-square of the simulated field over all frequency bins and all nodes of a floor plan, so the SNR refers to the field a microphone observes. The noise is white across bins in the whitened domain the model works in. A physical microphone noise floor is white in the time domain before whitening, so this is a controlled perturbation and not a calibrated model of microphone noise. One realization is drawn for each floor plan and source position, and every method sees the same realization. Table 5 compares all three reconstructors with and without noise.

Table 5: **Off-microphone NMSE (dB) with and without measurement noise**, 50 unseen floor plans; lower is better.

	Microphones (K)	4	16	64	128
Noise-free	Textbook kernel interpolation	-0.24	-0.86	-2.46	-3.39
	Matched kernel interpolation	-1.01	-1.66	-3.23	-4.12
	Ours	$\mathbf{-2.58}$	$\mathbf{-3.43}$	$\mathbf{-6.39}$	$\mathbf{-9.62}$
30 dB SNR	Textbook kernel interpolation	-0.24	-0.86	-2.46	-3.39
	Matched kernel interpolation	-1.01	-1.66	-3.23	-4.12
	Ours	$\mathbf{-2.59}$	$\mathbf{-3.43}$	$\mathbf{-6.38}$	$\mathbf{-9.60}$
20 dB SNR	Textbook kernel interpolation	-0.24	-0.85	-2.44	-3.38
	Matched kernel interpolation	-1.01	-1.65	-3.21	-4.10
	Ours	$\mathbf{-2.59}$	$\mathbf{-3.41}$	$\mathbf{-6.32}$	$\mathbf{-9.44}$
10 dB SNR	Textbook kernel interpolation	-0.20	-0.77	-2.31	-3.25
	Matched kernel interpolation	-0.97	-1.55	-3.05	-3.96
	Ours	$\mathbf{-2.56}$	$\mathbf{-3.28}$	$\mathbf{-5.71}$	$\mathbf{-8.12}$

The primary statistic, specified before the experiment was run, is the change in the margin over matched kernel interpolation at $K{=}64$, paired within each floor plan across the noise-free and noisy conditions. At 20 dB SNR it is $+0.06 \pm 0.02$ dB. The margin narrows slightly under noise, and our method remains more accurate than matched interpolation on all 50 floor plans at every budget. At 10 dB SNR the margin narrows by 0.50 ± 0.10 dB with 64 microphones and by 1.34 ± 0.23 dB with 128. Our error rises by 0.68 and 1.50 dB at these budgets, and that of matched interpolation by 0.18 and 0.16 dB, so the prior is the more sensitive of the two to noise at large budgets. Our method still outperforms matched interpolation on all 50 floor plans at every budget.

F.3 Training seeds

We trained our prior with two further seeds and every other setting unchanged, and evaluated each model exactly as the reported one. Table 6 gives the margin over matched kernel interpolation with 128 microphones. The three margins span 0.46 dB, and the model reported throughout the paper, seed 0, is the least accurate of the three.

Table 6: **Training seeds.** Our method minus matched kernel interpolation in off-microphone NMSE (dB, negative favors ours) with 128 microphones, and the number of floor plans on which ours is more accurate; $n{=}50$.

Seed	Ours − matched interpolation	Floor plans won
0 (reported)	-5.51 ± 0.46	50/50
1	-5.74 ± 0.47	50/50
2	-5.97 ± 0.50	50/50

F.4 Floor-plan errors

Our method and matched kernel interpolation both take the floor plan as given. To measure the cost of an inaccurate floor plan, we perturb the floor plan given to each method and recompute the ISM field and the interior mask from it, while the simulated ground truth, the lattice and the microphones stay unchanged. The first perturbation moves every wall outward by the same distance, 5, 10 or 20 cm. The second moves each wall segment independently along its normal by a Gaussian offset with a standard deviation of 5 cm. A draw that makes the floor plan self-intersect is redrawn, which was needed for 24 of the 50 floor plans. Table 7 reports the cost with 64 microphones and one source position per floor plan.

Table 7: **Floor-plan errors.** Increase in off-microphone NMSE (dB, positive is worse) when the floor plan given to the method is perturbed, for our method and for matched kernel interpolation, and our method minus matched interpolation under the same perturbation, with the number of floor plans on which ours is more accurate. The last row moves each wall independently. 64 microphones, one source position, $n{=}50$.

Perturbation	Ours	Matched	Ours − matched	Floor plans won
All walls outward, 5 cm	$+1.70 \pm 0.18$	$+0.42 \pm 0.05$	-1.87 ± 0.19	50/50
All walls outward, 10 cm	$+3.19 \pm 0.28$	$+0.64 \pm 0.08$	-0.60 ± 0.11	48/50
All walls outward, 20 cm	$+4.10 \pm 0.35$	$+0.72 \pm 0.09$	$+0.24 \pm 0.12$	14/50
Each wall, $\sigma{=}5$ cm	$+0.81 \pm 0.18$	$+0.28 \pm 0.05$	-2.62 ± 0.24	50/50

Our method relies on the ISM field more than matched interpolation does, so the same error costs it more. A uniform offset of 5 cm costs our method 1.70 dB and matched interpolation 0.42 dB, and at 20 cm matched interpolation is the more accurate of the two. Independent errors with the same standard deviation cost our method about half as much, 0.81 ± 0.18 dB, and it remains more accurate than matched interpolation on every floor plan. On the 26 floor plans whose first draw was valid, the costs are 0.88 ± 0.27 dB for our method and 0.29 ± 0.08 dB for matched interpolation.

F.5 Acquisition gain at every budget

Table 8 gives the improvement of A-optimal placement over farthest-point placement at every budget of the comparison in Section 4.3, with batches of $p=6$. At $K=4$ both rules use the same four initial microphones, so the difference there is zero by construction.

Table 8: **Improvement of A-optimal over farthest-point placement** in off-microphone NMSE (dB, positive favors A-optimal), with the number of floor plans on which A-optimal placement is more accurate; $n=50$, batches of $p=6$.

K	10	16	22	28	34
Improvement	0.26 ± 0.04	0.39 ± 0.06	0.47 ± 0.06	0.55 ± 0.07	0.59 ± 0.07
Floor plans won	50/50	50/50	50/50	50/50	50/50
K	40	46	52	58	64
Improvement	0.59 ± 0.08	0.61 ± 0.09	0.62 ± 0.10	0.65 ± 0.11	0.67 ± 0.11
Floor plans won	50/50	49/50	49/50	49/50	48/50

Interpolating each floor plan's farthest-point curve in $\log K$, and using the farthest-point result at $K=128$ above the last rung, farthest-point placement needs 75.41 ± 2.29 microphones to match A-optimal placement with 64. The two rules also spend their budgets in different places. We split each floor plan at the median, over its nodes, of the residual energy with respect to the ISM field, so that the lower half is where the ISM field alone is already most accurate. Farthest-point placement puts $40.1 \pm 1.7\%$ of its 64 microphones in that half and A-optimal placement puts $24.9 \pm 1.7\%$ there, a paired difference of 15.2 ± 2.2 percentage points in favor of the harder half, on 48 of 50 floor plans.

F.6 Cross-reconstructor replay

To test whether A-optimal placement wins only because the same posterior selects the microphones and reconstructs from them, we reconstruct from the same selected positions with matched kernel interpolation, which plays no part in selecting them. Matched interpolation also prefers the A-optimal positions, by 0.34 ± 0.05 dB at both $K=16$ and $K=64$, on 49 of 50 floor plans at each. Because both reconstructors see the same selections on the same floor plans, the part of the gain that depends on our model can be measured directly as a paired difference-in-differences, (ours, A-optimal − ours, farthest-point) − (interpolation, A-optimal − interpolation, farthest-point). It is 0.05 ± 0.05 dB at $K=16$ and 0.33 ± 0.09 dB at $K=64$. With 16 microphones, 86% of the gain transfers to interpolation and comes from the positions themselves; with 64, 51% transfers, and our model uses the positions better than interpolation does.

F.7 Measurement efficiency

We also express each gain as a number of microphones. Figure 6 evaluates matched kernel interpolation at 44 budgets from 4 to 512 microphones, on farthest-point designs and the protocol of Table 1; each design extends the design with fewer microphones. For a floor plan, a target NMSE and a reference method, K' is the number of microphones at which the reference first reaches the target, interpolated linearly in $\log K$ between measured budgets. When the reference does not reach the target with 512 microphones, we record $K'=512$ as a lower bound. Medians carry percentile bootstrap intervals over floor plans (2,000 resamples), and means carry t intervals.

For the prior, the target is our NMSE with K microphones and the reference is matched interpolation. Table 9 gives the median ratio K'/K. Matched interpolation needs 10 times as many microphones as our method with 4, and 4.6 times as many as our method with 16. With 64 and 128 microphones, it does not reach our accuracy with 512 on 49 and 50 of the 50 floor plans, so the ratios there are lower bounds. The smallest median ratio at 4, 16 and 64 microphones is 4.6.

For active sampling, the target is A-optimal placement with K microphones, in batches of $p=6$ with one source position per floor plan, and the reference is our method with farthest-point placement at the budgets of Table 8 and at 128 microphones. Table 10 gives the mean K' and the saving $1 - K/\bar{K}'$, where $\bar{K}'$ is the mean of K' over floor plans. With 64 microphones the saving is 15%, and it is larger at smaller budgets.

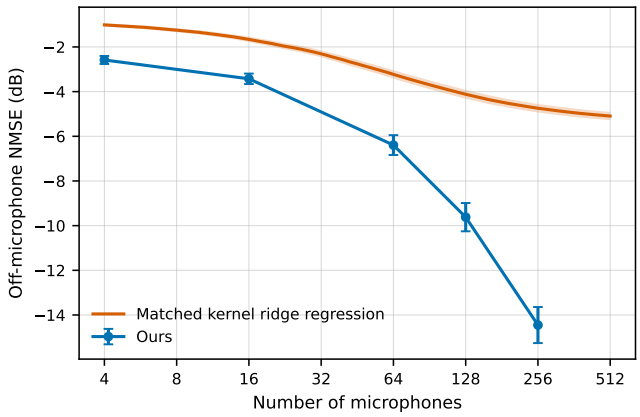


Figure 6: **Error against the number of microphones.** Matched kernel interpolation at 44 budgets and our method at the budgets of Table 1, mean and 95% CI over the 50 unseen floor plans.

Table 9: **Microphones matched interpolation needs to reach our accuracy.** Median over floor plans of K'/K with its bootstrap 95% CI, the median K', and the number of floor plans on which matched interpolation does not reach our NMSE with 512 microphones; n=50.

Our budget K	Median K'/K	Median K'	Not reached by 512
4	10.25 [9.21, 11.19]	41	0/50
16	4.63 [4.35, 4.87]	74	0/50
64	≥ 8	≥ 512	49/50
128	≥ 4	≥ 512	50/50

For the replay of Appendix F.6, the target is matched interpolation on the A-optimal positions with 64 microphones, and the reference is matched interpolation on farthest-point designs, both with one source position. The reference reaches the target with a median of 81.1 microphones (95% CI 77.9 to 84.8), a saving of 21%.

F.8 Guidance weight, sample count and step count

The guidance weight ζ was set once to 10 on training floor plans and kept fixed for every result. Table 11 re-runs reconstruction with other weights and everything else fixed: the same checkpoint, microphones, posterior samples, step count and seed.

Every weight on either side of ζ=10 is less accurate at every budget tested, so 10 is an interior optimum among the values tested. Small weights are the failure mode of Section 3.2: too weak a correction leaves the samples close to the prior's unconditional field.

On 9 unseen floor plans we also vary the number of integration steps T and posterior samples S around T=20 and S=8. Halving the steps to T=10 costs 0.62 ± 0.24 dB at K=64, and doubling them to T=40 changes NMSE by -0.12 ± 0.11 dB, which is within noise. More samples are more accurate on all 9 floor plans: S=16 by 0.14 ± 0.02 dB and S=32 by 0.20 ± 0.03 dB. Both gains are below 0.5 dB, and S=8 trades them for a four-fold saving in sampling cost.

F.9 Choice of sampler

Table 2 replaces guided posterior sampling with a projection sampler, DDNM, which enforces each measurement at its own node, with the prior, microphones and budget unchanged. With 64 microphones, the projection sampler is less accurate by 3.66 ± 0.31 dB, on every unseen floor plan. Splitting the error by location shows why. On the first 10 unseen floor plans, the projection sampler reaches -104.59 dB at the microphones, the floating-point floor, and -2.52 dB away from them, while guided sampling reaches -16.92 dB at the microphones and -5.96 dB away from them. For

Table 10: **Farthest-point microphones needed to match A-optimal placement** with our method: mean $\bar{K}'$ with its 95% CI, and the saving $1 - K/\bar{K}'$ with the interval implied by that CI; n=50, one source position, batches of p=6.

A-optimal budget K	10	16	22	28	34
Mean K'	13.90 ± 0.66	22.67 ± 1.15	30.44 ± 1.25	37.45 ± 1.46	44.41 ± 1.78
Saving (%)	28.1	29.4	27.7	25.2	23.4
95% CI	$[24.5, 31.3]$	$[25.7, 32.8]$	$[24.6, 30.6]$	$[22.2, 28.0]$	$[20.2, 26.4]$
A-optimal budget K	40	46	52	58	64
Mean K'	51.16 ± 2.04	57.71 ± 2.02	63.85 ± 1.94	69.69 ± 2.14	75.41 ± 2.29
Saving (%)	21.8	20.3	18.6	16.8	15.1
95% CI	$[18.6, 24.8]$	$[17.4, 23.0]$	$[16.0, 21.0]$	$[14.1, 19.2]$	$[12.5, 17.6]$

Table 11: **Cost of other guidance weights**, relative to ζ=10, in off-microphone NMSE (dB, positive is worse). ζ=0.3 uses all 50 unseen floor plans; the other weights use 24. No floor plan is more accurate at any weight other than ζ=10.

Guidance weight	K=16	K=64	K=128
ζ=0.3 (n=50)	$+0.89 \pm 0.09$	$+3.69 \pm 0.32$	—
ζ=1 (n=24)	$+0.49 \pm 0.08$	$+2.73 \pm 0.36$	$+5.44 \pm 0.60$
ζ=3 (n=24)	$+0.11 \pm 0.03$	$+1.15 \pm 0.17$	$+2.76 \pm 0.31$
ζ=30 (n=24)	$+1.07 \pm 0.17$	$+0.71 \pm 0.12$	$+0.70 \pm 0.16$

a point-sampling operator, a projection step changes the field only at the measured nodes and leaves every other node to the unconditional prior. Guidance passes the same misfit through the network's Jacobian, which carries it to every node.

F.10 Conditioning-field support

At evaluation the ISM field is supported on each arrangement's nodes, so it has gaps where furniture stands, whereas in training it covers the bare room; the interior mask is the bare room in both. Re-running our method with the field on the bare room's support changes the mean NMSE by at most 0.01 dB at any budget from 4 to 128 microphones (n=50).