HEARING ANYWHERE FROM ANYTHING: GEOMETRY-AWARE ACOUSTIC FIELD SYNTHESIS FROM UNLOCALIZED BLIND RECORDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing a room's acoustic field is essential for rendering spatial audio at previously unobserved source locations. Yet existing methods typically require either scene-specific optimization or carefully measured room impulse responses (RIRs) with calibrated positions. We observe that ordinary recordings in everyday life, such as speech and music, are also shaped by the room and therefore contain acoustic evidence. Exploiting this evidence is challenging because both the dry source signals and the positions of the recorded sources are unknown. Our key insight is to formulate it as a physics-anchored conditional completion problem. Unlocalized recordings reveal room-level decay and reverberation, whereas target geometry anchors direct-path timing and specifies the source–receiver path of interest. Building on this insight, we introduce PHILAC, a two-stage framework that first predicts a query-conditioned energy-decay profile and then generates the target RIR through masked discrete diffusion of acoustic codec tokens. On unseen synthetic rooms, PHILAC achieves a mean EDC error of 2.660 dB, improving over the strongest baseline by 19.5%. On real rooms, we outperform the strongest baseline by 45.7% in mean EDC error without scene-specific adaptation. These results show that everyday recordings can provide acoustic context for synthesizing RIRs at new locations without calibrated acoustic measurements.

1 INTRODUCTION

Rendering spatial audio in a real environment requires knowing how sound propagates between arbitrary positions. This propagation is characterized by the room impulse response (RIR), which captures the direct path, reflections, and reverberation between them. Reconstructing a room's acoustic field can therefore be formulated as recovering a spatially queryable field of RIRs. Such a representation enables acoustically consistent rendering of novel or moving sources, supporting applications in augmented and virtual reality (Chen et al., 2022; Tong et al., 2025), video re-rendering (Li et al., 2025; Yoshida et al., 2026; Kim et al., 2025; Liu et al., 2025a), robotics (Xu et al., 2026; Xie et al., 2026), and acoustic digital twins (Lan et al., 2026). In practice, however, constructing such a field requires substantial measurement effort, involving controlled acoustic excitation and measurements over many source–receiver configurations.

Existing methods have progressively reduced the number of measurements required for acoustic-field reconstruction. Neural acoustic fields interpolate densely sampled RIRs through scene-specific optimization (Luo et al., 2022; Wang et al., 2024; Su et al., 2022), while recent cross-room methods leverage deterministic or generative priors to reconstruct the field from only a few observations (Brunetto, 2026; Jing et al., 2026). Despite this progress, these methods still rely on specially measured RIRs as acoustic context, along with calibrated source–receiver locations of each measurement. Obtaining even a sparse set of such calibrated context requires active acoustic probing and meticulous spatial registration in the target environment, limiting their deployment in spontaneous, uncalibrated scenarios.

To bypass these constraints, a natural alternative is to infer acoustic context from sounds that already occur in the environment. Ordinary recordings, such as speech and music, carry information about the room through the propagation paths. Prior work has demonstrated that such ordinary recordings

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.

Table 1: Overview of related work. Measured: probe RIR or controlled excitation. Blind: ordinary recordings, no probe RIR or context pose. Query: Observed path vs. a different Unobserved path. Transfer: Cross-room and real-room.

Method	Observation	Query	Geometry	Transfer	Generative	Microphone
NAF (Luo et al., 2022)	Measured	Unobserved	-	Real	-	Stereo / Monaural
INRAS (Su et al., 2022)	Measured	Unobserved	✓	Real	-	Monaural
NACF (Liang et al., 2023b)	Measured	Unobserved	✓	-	-	Monaural
AV-NeRF (Liang et al., 2023a)	-	Unobserved	✓	Real	-	Binaural
DiffRIR / HAA (Wang et al., 2024)	Measured	Unobserved	✓	Real	-	Monaural / Binaural
NeRAF (Brunetto et al., 2025)	Measured	Unobserved	✓	Real	-	Stereo
FewShotRIR (Majumder et al., 2022)	Measured	Unobserved	✓	Cross	-	Binaural
MAGIC (Huang et al., 2024)	Measured	Unobserved	✓	Cross	-	Binaural
xRIR (Liu et al., 2025b)	Measured	Unobserved	✓	Cross, Real	-	Monaural
FLAC (Brunetto, 2026)	Measured	Unobserved	✓	Cross, Real	✓	Monaural
EigeNet (Jing et al., 2026)	Measured	Unobserved	✓	Cross, Real	-	Monaural
Speech2IR (Ratnarajah et al., 2023)	Blind	Observed	-	Cross, Real	-	Monaural
AV-RIR (Ratnarajah et al., 2024)	Blind	Observed	✓	Cross, Real	-	Monaural
BUDDy (Moliner et al., 2024)	Blind	Observed	-	Cross, Real	✓	Monaural
Blind SRIR (Lluís, 2025)	Blind	Unobserved	-	Cross	✓	Tetrahedral
DARAS (Wang et al., 2025a)	Blind	Observed	-	Cross, Real	-	Monaural
MEAN-RIR (Chen et al., 2025)	Blind	Observed	✓	Cross	-	Monaural
AnyRIR (Lee et al., 2026)	Blind	Observed	-	Real	-	Monaural
Gencho (Lin et al., 2026)	Blind	Observed	-	Cross, Real	✓	Monaural
PHILAC (Ours)	Blind	**Unobserved**	✓	Cross, Real	✓	Monaural

can reveal global acoustic attributes (Eaton et al., 2016; Wang et al., 2025b) or translate an observed recording into its underlying RIR (Lee et al., 2026; Wang et al., 2025a; Ratnarajah et al., 2023). Yet, both capabilities remain inadequate for full acoustic field reconstruction. Global acoustic attributes offer no spatially varying fine structure, while single-instance RIR estimation cannot extrapolate to unobserved source locations. Bridging this gap requires a paradigm shift, *i.e.*, extracting transferable acoustic evidence from uncalibrated recordings and synthesizing RIRs at arbitrary, unobserved target positions. This motivates our research problem: *can a room's acoustic field be reconstructed from a few ordinary recordings and generalized to previously unobserved source locations?*

Addressing this question introduces two fundamental challenges that depart from conventional RIR-based field reconstruction. First, ordinary recordings entangle environmental reverberation with unknown, time-varying source content, where dry signal dynamics often overshadow subtle room cues. Second, the exact poses of the context sound sources are completely unknown, depriving the model of explicit spatial propagation references.

Our insight is to formulate blind-context acoustic field reconstruction as a *physics-anchored conditional completion problem*. The core intuition is that query geometry natively determines direct-path timing, while uncalibrated recordings constrain room-level decay and reverberation. We pool this evidence across the unordered recordings without assigning them spatial coordinates, and combine it with the target geometry to plan the coarse energy decay of the requested response. A discrete RIR generator then completes the reflection structure and waveform details that neither input determines uniquely.

Based on this insight, we introduce PHILAC, the first discrete diffusion framework for blind-context acoustic field reconstruction. At a high level, PHILAC disentangles macro-level energy decay from micro-level waveform generation through a two-stage architecture: 1) **Acoustic Planner:** It combines position-free room evidence with target geometry to predict a query-level energy decay profile. 2) **Neural Renderer:** We tokenize RIRs using a discrete acoustic codec and formulate token field generation as a masked discrete diffusion process. Together, this architecture frees acoustic field reconstruction from mandatory positional registration of context sounds, achieving physics-grounded spatial generation with the efficiency required for practical applications.

We extensively evaluate PHILAC across diverse synthetic and real-world acoustic environments. Our method achieves a mean EDC error of 2.660 dB, outperforming the strongest competing baseline by 19.5% and establishing a new state of the art. Furthermore, under zero-shot transfer to real-world rooms, PHILAC maintains a low mean EDC error of 2.797 dB, representing a 45.7% error reduction. These results demonstrate highly accurate synthesis at unobserved source locations as well as strong sim-to-real transferability.

In summary, our contributions are as follows: ❶ We introduce a novel task of geometry-aware blind acoustic field reconstruction. Given ordinary audio recordings (*e.g.*, speech or music), our objective is to synthesize the RIR at an unobserved source location without requiring dry signals, calibrated source positions, or reference RIRs from the target environment. ❷ To our knowledge, we are the first to apply discrete diffusion to blind RIR synthesis. We propose PHILAC, a novel network that aggregates room cues from unregistered recordings and combines them with panoramic geometry and the target source query. ❸ We demonstrate state-of-the-art reconstruction fidelity and generation efficiency across diverse real-world and synthetic environments.

2 RELATED WORK

We present a summary of related work as shown in Tab. 1.

Neural acoustic fields. Neural acoustic fields model RIRs as continuous functions of source and receiver positions, enabling RIR synthesis at unmeasured locations (Luo et al., 2022; Su et al., 2022; Liang et al., 2023b). However, this paradigm typically relies on dense measurements across diverse source–receiver configurations and requires per-room optimization. While geometry-aware variants reduce the sampling burden (Brunetto et al., 2025), adapting to an unseen environment still demands scene-specific acoustic evidence. Consequently, strong intra-room interpolation capabilities do not easily generalize to cross-room prediction from a few uncalibrated recordings.

Few-shot acoustic field synthesis. Few-shot acoustic synthesis reduces the need for dense measurements by learning cross-room acoustic priors from a small set of observations. To infer unobserved acoustic paths, recent methods leverage visual and geometric cues to anchor observed RIRs or echoes (Majumder et al., 2022; Huang et al., 2024; Wang et al., 2024; Liu et al., 2025b; Jing et al., 2026). Moreover, generative models account for acoustic ambiguity when conditioned on sparse inputs (Brunetto, 2026). Nevertheless, these paradigms still suffer from an implicit measurement bottleneck where acquiring a single reference response requires explicit acoustic measurement and precise geometric calibration. PHILAC removes these requirements by learning from ordinary recordings whose dry signals and source positions are unknown.

Blind acoustic estimation. Existing blind acoustic estimation methods predominantly focus on either global acoustic parameters (Eaton et al., 2016; Wang et al., 2025b) or the specific RIR associated with an observed recording (Ratnarajah et al., 2023; 2024; Moliner et al., 2024; Wang et al., 2025a; Chen et al., 2025; Lin et al., 2026). However, recovering the RIR of an observed path reduces to a standard blind deconvolution task, whereas we target response prediction at completely unobserved positions. While Blind SRIR (Lluís, 2025) targets novel-path synthesis, it requires a tetrahedral microphone array within synthetic data while omitting scene geometry. PHILAC overcomes these restrictions by conditioning on separate monaural recordings paired with receiver-centered depth geometry, demonstrating successful generation in real rooms.

Discrete generative modeling. Discrete diffusion extends denoising to categorical sequences, including processes with an absorbing mask state (Austin et al., 2021). Masked diffusion language models have developed this formulation into effective sequence generators at increasing scale (Sahoo et al., 2024; Nie et al., 2026). Iterative masked prediction has also enabled parallel generation of image and audio codec tokens (Arellano et al., 2025; Borsos et al., 2023; Ziv et al., 2024). For RIRs, codec-based generators have used supplied geometric or acoustic parameters (Lyu et al., 2025; Arellano et al., 2025). PHILAC instead infers acoustic evidence from unlocated recordings and combines it with target geometry and a supervised decay plan to generate RIR tokens.

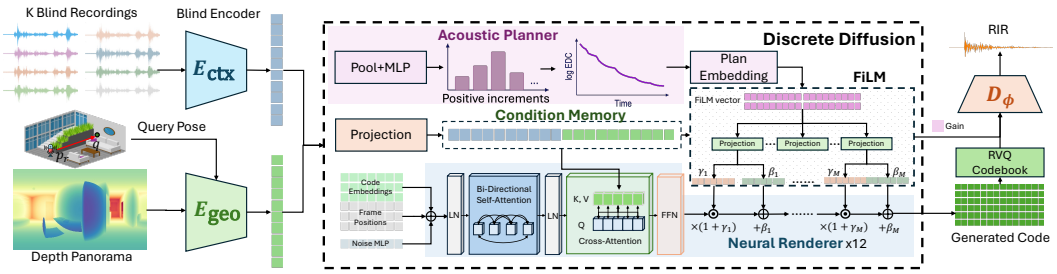


Figure 1: Overview of the PHILAC framework.

3 METHOD

3.1 PROBLEM FORMULATION

Let R denote a room with a fixed receiver at p_r. We observe an unordered set of K ordinary recordings $\mathcal{Y} = \{y_k\}_{k=1}^{K}$, where

$$y_k = s_k * h_R(p_k, p_r) + \epsilon_k. \tag{1}$$

Here s_k is the dry source signal, p_k is its position, $h_R(p_k, p_r)$ is the corresponding room impulse response (RIR), and ϵ_k is recording noise. Given $\mathcal{Y}$, a receiver-centered depth panorama G, and a known target source coordinate q relative to the receiver, our goal is to generate the target RIR $h_q = h_R(q, p_r)$.

This task is blind *twice*. First, the dry signal s_k is unknown: each recording entangles source content with room acoustics. Second, the context-source position p_k is unknown: the propagation paths represented by the recordings have no spatial registration. The recordings therefore cannot be treated as positioned RIR samples for interpolation. The target query is different. Its coordinate q and the receiver geometry G are available, even though the source positions of the context recordings are not. PHILAC leverages the recordings to infer room evidence shared across unknown sources and paths, then employs the target geometry to specify the path for which an RIR should be generated. It requires neither measured context RIRs nor a fragment of the target RIR.

3.2 PHYSICS-ANCHORED CONDITIONAL COMPLETION

We present our overall framework in Fig. 1. Generally, the recordings provide acoustic evidence about the room, while the known target geometry provides a nominal propagation-time anchor and describes the requested source–receiver path. A planner combines these conditions to predict the path's energy decay. A renderer then generates the waveform structure consistent with the available evidence and the predicted decay:

$$\begin{aligned} U &= E_{\mathrm{ctx}}(\mathcal{Y}), & V_q &= E_{\mathrm{geo}}(G, q), \\ d_q &= P_\psi(U, V_q), & (Z_q, z_q^g) &\sim R_\theta(\cdot \mid U, V_q, d_q), \\ \widehat{h}_q &= D_\phi(Z_q, z_q^g, \tau_q), & \tau_q &= \|q\|/c. \end{aligned} \tag{2}$$

Here U represents acoustic evidence aggregated from the unlocated recordings, V_q represents the target geometry, and d_q is a query-conditioned decay plan. The renderer predicts discrete tokens Z_q for the normalized response and a separate gain token z_q^g. A pretrained, frozen codec decodes these tokens, and the geometry-derived delay τ_q places the response in time.

This decomposition determines the order of the method: we first construct the two sources of conditioning, then predict the target decay, and finally generate the waveform detail that the decay plan alone cannot specify.

3.3 BLIND CONTEXT ENCODING AND QUERY-RELATIVE GEOMETRY

The two conditioning branches address different parts of the problem. The audio branch extracts useful room cues without knowing which source or path produced each recording. The geometry

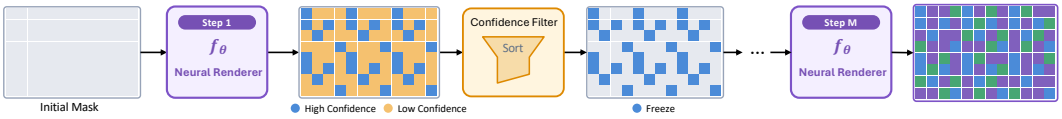


Figure 2: **Confidence-Guided Masked Decoding of RIR Tokens.** Generation starts from a fully masked token grid. At each iteration, the renderer proposes tokens in parallel under the room, query-geometry, and decay-plan conditions. Confident predictions are retained while uncertain positions are revisited, until the completed grid is decoded into the target RIR.

branch describes the particular target path without assigning a position to any context recording. For the audio recordings, we compute the log-STFT features and process them with a shared audio encoder, yielding a sequence of tokens X_k.

Because the recordings have no known spatial roles, we process them as a set. Learned room queries attend to the union of their encoded features via a Set Transformer-style attention (Lee et al., 2019):

$$U = \mathrm{SetAttn}\left(Q_R, \bigcup_{k=1}^{K} X_k\right). \tag{3}$$

Neither recording-order embeddings nor context-position embeddings are used, making the aggregation invariant to the order of the recordings. Multiple recordings expose the model to different source signals and propagation paths in the same room; the resulting tokens can retain acoustic evidence that recurs across those observations. An auxiliary probabilistic head predicts room-acoustic descriptors from pooled tokens to encourage this evidence to remain accessible in U.

The aggregated tokens describe what the recordings suggest about the room, but they do not identify the path requested by q. For that purpose, the geometry branch expresses the visible surfaces relative to the target source. Let $d_G(\omega)$ be the depth in panorama direction ω, $r(\omega)$ its unit ray, and $x_G(\omega) = d_G(\omega)r(\omega)$ the observed surface point in the receiver frame. We form the query-relative surface field

$$M_q(\omega) = q - x_G(\omega). \tag{4}$$

A geometry encoder E_{geo} processes this field and appends a query token derived from the target coordinate, distance, and nominal propagation delay:

$$V_q = E_{\mathrm{geo}}(M_q, q, \|q\|, \tau_q). \tag{5}$$

Thus, U carries evidence from the unlocated recordings, while V_q specifies where that evidence must be applied. Both enter the planner and the renderer; the target geometry does not depend on recovering the unknown context-source positions.

3.4 Query-Conditioned Decay Planning

Room evidence pooled across unknown paths cannot by itself specify the energy evolution of a particular target RIR. The decay planner bridges this gap by combining U and V_q to predict a query-conditioned energy-decay curve. It predicts positive increments rather than independent curve values:

$$\boldsymbol{\delta}_q = \mathrm{softplus}\big(f_\psi\big(\mathrm{Pool}(U), \mathrm{Pool}(V_q)\big)\big), \tag{6}$$

where $\mathrm{softplus}(x) = \log(1 + e^x)$. The normalized log-energy curve is then constructed as

$$m_{q,b} = m_{q,b-1} - \delta_{q,b}, \qquad m_{q,0} = 0, \tag{7}$$

so it is monotone by construction. A differentiable fit to the curve produces a reverberation-time estimate $\widehat{T}_q$. The curve and the fitted time are embedded into the plan via a learned MLP, *i.e.*,

$$d_q = \mathrm{PlanEmbed}(m_q, \widehat{T}_q). \tag{8}$$

During training, the target RIR supplies EDC and reverberation-time supervision. At inference, the planner predicts both from (U, V_q). We stop renderer gradients at d_q so that the plan remains tied to these supervised acoustic quantities. The plan guides the large-scale decay of the requested response, but an energy-decay curve does not determine individual reflections or waveform phase. Those details remain a generative completion problem.

3.5 TIME- AND GAIN-FACTORED RIR TOKENIZATION

We represent waveform detail with a neural RIR codec pretrained before the conditional model. Its residual vector quantizer (RVQ) maps a response to a grid of discrete tokens. Two other properties are handled separately: the nominal propagation delay is available from the target coordinate, and overall gain is represented by its own token. The waveform grid can consequently focus on the shape of an aligned, normalized response.

Before encoding, we align it to the canonical anchor using the geometry-derived delay. Let h_q^{can} denote the aligned response. We then separate its peak gain g_q from the normalized waveform:

$$g_q = \|h_q^{\mathrm{can}}\|_\infty, \qquad Z_q = Q_\phi\big(E_\phi(h_q^{\mathrm{can}}/g_q)\big), \qquad z_q^g = Q_g(\log g_q). \tag{9}$$

Here E_ϕ is a causal waveform encoder, Q_ϕ is the RVQ module, and Q_g is a scalar gain quantizer. The residual codebooks capture progressively finer response structure. Given predicted tokens and gain, decoding restores the canonical waveform and then its target delay. The alignment and decoding operations are given in Appendix A.2.

After pretraining, the codec encoder, quantizers, and decoder are frozen. The encoder supplies target tokens during conditional-model training; the codebooks and decoder turn predicted tokens back into an RIR at inference. The remaining task is therefore to predict (Z_q, z_q^g) from the blind recordings and the target query.

3.6 DECAY-GUIDED MASKED RIR RENDERING

We formulate token prediction as absorbing-state discrete diffusion. For token z_0^j at a joint frame–codebook position j, a forward process replaces it with a mask according to schedule $\rho(t)$:

$$q(z_t^j \mid z_0^j) = \big(1-\rho(t)\big)\mathbf{1}[z_t^j = z_0^j] + \rho(t)\mathbf{1}[z_t^j = \texttt{[MASK]}]. \tag{10}$$

A bidirectional Transformer observes the corrupted grid and predicts the masked positions jointly. Its cross-attention memory combines the detailed tokens from both conditioning branches:

$$C_q = [U; V_q]. \tag{11}$$

This memory lets each prediction access evidence from the recordings and the query-relative surfaces. The decay plan supplies a complementary, global condition through FiLM modulation (Perez et al., 2018) in every Transformer block (Vaswani et al., 2017):

$$\widetilde{X}_\ell = T_\ell(X_{\ell-1}, C_q), \qquad X_\ell = \big(1+\gamma_\ell(d_q)\big) \odot \widetilde{X}_\ell + \beta_\ell(d_q). \tag{12}$$

Separate categorical heads predict the RVQ codebooks, while a condition head predicts the gain token. In this way, the memory provides detailed room and path information, and the supervised plan guides the response's overall decay during token generation. The plan constrains the renderer through learned conditioning; it does not uniquely determine the generated waveform.

3.7 LEARNING AND CONFIDENCE-GUIDED DECODING

The frozen codec supplies target RVQ and gain tokens for each training RIR. We train the renderer to recover masked tokens conditioned on the blind recordings, target geometry, and predicted decay. An additional fully masked pass trains the renderer to make its initial prediction without visible target tokens. The objective also supervises gain, room descriptors, and the decay planner. Appendix A.3 gives the complete objective.

At inference, the context encoder processes $\mathcal{Y}$ once to obtain U, which can be reused for multiple target queries. For each q, the geometry encoder and planner produce V_q and d_q. As illustrated in Figure 2, decoding begins with a fully masked token grid. At each iteration, the renderer proposes tokens for the grid in parallel, retains confident predictions, and revisits uncertain positions. The completed grid and predicted gain are decoded by the frozen codec; the geometric shift then places the response at its nominal propagation time. The resulting RIR is conditioned on the observed room and the requested path without estimating context-source positions or observing any part of the target response.

Table 2: Performance on unseen AcousticRooms scenes with $K = 8$. Entries are query mean $\pm$ std.

Method	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
Nearest	$5.999_{\pm 4.083}$	$7.402_{\pm 5.845}$	$41.43_{\pm 41.89}$	$3.980_{\pm 3.681}$	$0.2107_{\pm 0.2698}$	$9.726_{\pm 3.803}$
Direct interpolation	$7.658_{\pm 4.422}$	$8.608_{\pm 6.178}$	$36.39_{\pm 33.24}$	$8.167_{\pm 5.017}$	$0.4469_{\pm 0.3211}$	$16.374_{\pm 3.820}$
Blind T60+FDN (Jot & Chaigne, 1991)	$5.795_{\pm 3.900}$	$7.260_{\pm 5.870}$	$35.46_{\pm 26.33}$	$3.985_{\pm 3.702}$	$0.2126_{\pm 0.2814}$	$9.726_{\pm 3.803}$
FLAC (Brunetto, 2026)	$3.305_{\pm 1.844}$	$4.055_{\pm 2.250}$	$21.98_{\pm 16.57}$	$2.508_{\pm 3.058}$	$0.1043_{\pm 0.1009}$	$2.323_{\pm 1.692}$
EigeNet (Jing et al., 2026)	$3.375_{\pm 1.989}$	$4.152_{\pm 2.520}$	$21.92_{\pm 16.77}$	$2.721_{\pm 4.754}$	$0.1256_{\pm 0.1773}$	$2.328_{\pm 1.948}$
xRIR (Liu et al., 2025b)	$6.922_{\pm 4.537}$	$8.149_{\pm 5.306}$	$42.22_{\pm 35.28}$	$6.056_{\pm 5.851}$	$0.2797_{\pm 0.2113}$	$6.499_{\pm 3.440}$
Speech2RIR (Ratnarajah et al., 2023)	$23.185_{\pm 2.956}$	$32.129_{\pm 11.542}$	$88.34_{\pm 76.08}$	$5.774_{\pm 5.809}$	$0.4569_{\pm 0.2088}$	$3.357_{\pm 2.669}$
Blind SRIR (Lluís, 2025)	$6.437_{\pm 4.509}$	$7.570_{\pm 4.917}$	$89.45_{\pm 96.76}$	$3.310_{\pm 2.915}$	$0.2033_{\pm 0.1723}$	$3.130_{\pm 2.187}$
PHILAC	$2.660_{\pm 1.848}$	$3.216_{\pm 2.276}$	$18.39_{\pm 14.24}$	$2.031_{\pm 1.617}$	$0.0957_{\pm 0.0913}$	$1.831_{\pm 1.592}$
PHILAC + blind selection	$\mathbf{2.543}_{\pm \mathbf{1.783}}$	$\mathbf{3.095}_{\pm \mathbf{2.210}}$	$\mathbf{17.56}_{\pm \mathbf{14.06}}$	$\mathbf{1.764}_{\pm \mathbf{1.637}}$	$\mathbf{0.0848}_{\pm \mathbf{0.0911}}$	$\mathbf{1.767}_{\pm \mathbf{1.562}}$

Table 3: Effect of the number of blind context recordings. Entries are query mean $\pm$ std.

K	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
1	$3.916_{\pm 2.992}$	$4.692_{\pm 3.528}$	$27.55_{\pm 27.14}$	$2.687_{\pm 2.040}$	$0.1405_{\pm 0.1353}$	$2.182_{\pm 1.788}$
2	$3.317_{\pm 2.406}$	$3.981_{\pm 2.875}$	$22.76_{\pm 19.20}$	$2.406_{\pm 1.841}$	$0.1181_{\pm 0.1111}$	$2.024_{\pm 1.696}$
4	$2.934_{\pm 2.045}$	$3.535_{\pm 2.486}$	$20.27_{\pm 16.23}$	$2.156_{\pm 1.679}$	$0.1033_{\pm 0.0981}$	$1.889_{\pm 1.602}$
8	$2.660_{\pm 1.848}$	$3.216_{\pm 2.276}$	$18.39_{\pm 14.24}$	$2.031_{\pm 1.617}$	$0.0957_{\pm 0.0913}$	$1.831_{\pm 1.592}$

4 EVALUATION

Implementations: We train PHILAC on the room-disjoint AcousticRooms split. We leverage LibriSpeech speech (Panayotov et al., 2015), MUSDB18-HQ music (Rafii et al., 2017; 2019), and broadband noise as the dry sources. A BERP-inspired recording encoder (Wang et al., 2025b) aggregates the unordered observations into 16 room tokens. We pretrain and freeze an RIR codec with eight 512-entry residual codebooks before training the conditional model. Further architecture, optimization, and baseline-adaptation details are given in Appendix B.

Evaluation Setting: We evaluate whether blind recordings can support RIR prediction at source locations that were never observed acoustically. Unless stated otherwise, we use an eight-shot setting($K = 8$), with eight ordinary recordings per room whose dry signals and source positions are unknown. For each room, the model receives ordinary recordings with unknown dry signals and unknown source positions, together with receiver-centered geometry and the target source coordinate. We first test on unseen AcousticRooms scenes, where changing the number and content of the recordings allows us to examine what information the model extracts. We then evaluate zero-shot transfer to real rooms. Reconstruction is assessed using full and late energy-decay curve (EDC) errors, relative T_{60} error, and errors in C_{50}, early decay time EDT, and direct-to-reverberant ratio (DRR). Details of the metrics are present in Appendix C.

4.1 PERFORMANCE ON UNSEEN ROOMS

Overall Results: As shown in Tab. 2, PHILAC achieves the lowest error on all six metrics among single-sample methods. Its EDC error of 2.660 dB is 19.5% lower than the SOTA result of 3.305 dB. These results demonstrate that PHILAC handles the blind RIR synthesis task.

Among few-shot RIR synthesis methods, FLAC and EigeNet are the strongest competitors, whereas xRIR reaches 6.922 dB EDC error. Nearest response selection, direct interpolation, and blind T_{60} estimation followed by an FDN yield errors between 5.795 and 7.658 dB. These results indicate that neither transferring a response from an unregistered observation nor specifying the room through a coarse decay parameter is sufficient to reconstruct the requested path.

Blind RIR estimators face a different task mismatch. Speech2RIR estimates the response associated with an observed recording, rather than a response at a separate query, and obtains 23.185 dB EDC

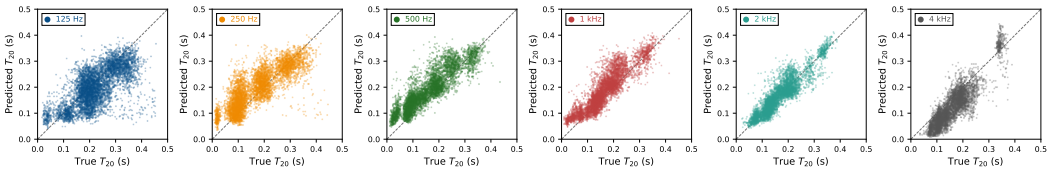


Figure 3: Predicted versus ground-truth octave-band T_{20}.

Table 4: Zero-shot transfer to real HAA rooms.

Method	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
FLAC (Brunetto, 2026)	$5.147_{\pm 3.251}$	$6.845_{\pm 4.728}$	$22.19_{\pm 22.38}$	$4.529_{\pm 3.599}$	$0.3975_{\pm 0.3785}$	$4.401_{\pm 2.799}$
EigeNet (Jing et al., 2026)	$6.375_{\pm 3.911}$	$8.737_{\pm 6.365}$	$28.54_{\pm 23.14}$	$5.237_{\pm 3.918}$	$0.4588_{\pm 0.4155}$	$\mathbf{3.993}_{\pm \mathbf{2.937}}$
xRIR (Liu et al., 2025b)	$6.848_{\pm 5.022}$	$9.094_{\pm 6.488}$	$46.92_{\pm 73.52}$	$5.978_{\pm 5.527}$	$0.4467_{\pm 0.3848}$	$5.833_{\pm 4.588}$
PHILAC	$\mathbf{2.797}_{\pm \mathbf{2.339}}$	$\mathbf{3.165}_{\pm \mathbf{2.649}}$	$\mathbf{17.97}_{\pm \mathbf{26.08}}$	$\mathbf{3.852}_{\pm \mathbf{2.115}}$	$\mathbf{0.1869}_{\pm \mathbf{0.1589}}$	$4.041_{\pm 2.521}$

error here. Blind SRIR addresses spatial extrapolation but uses different microphone and spatial-input assumptions; its error is 6.437 dB under the evaluated adaptation. These comparisons place the results in context without treating the methods' native operating conditions as identical.

Besides, we include evaluations of the blind selection. For the same query, we generate N candidate RIRs and select the one whose EDC best matches the planner's predicted EDC. It reduces EDC error from 2.660 to 2.543 dB, showing that the predicted decay can also guide the choice among generated responses. Because this variant requires multiple candidates, all subsequent analyses of the base model use one sample per query.

Number of recordings: We further evaluate how the number of blind recordings affects synthesis. As shown in Tab. 3, PHILAC's EDC error decreases monotonically from 3.916 to 3.317, 2.934, and 2.660 dB as K increases from 1 to 2, 4, and 8, respectively. This amounts to a 32.1% reduction from one to eight recordings. The other five metrics follow the same improving trend. The largest EDC gain occurs between one and two recordings, while subsequent recordings provide smaller gains. Since the source signals and positions remain unknown, these results indicate that PHILAC can accumulate useful acoustic evidence without spatially registering the observed paths.

Octave-Band Analysis: Full-band errors cannot determine whether a generated RIR also reproduces frequency-dependent acoustic behavior. We therefore examine octave-band reverberation-time predictions. In Fig. 3, predicted T_{20} generally follows the measured variation across bands, but the lower-frequency bands exhibit greater dispersion. Matching the overall decay is therefore easier than reproducing every frequency-dependent reverberation time.

4.2 ZERO-SHOT TRANSFER TO REAL ROOMS

We next evaluate generalization to measured rooms on the HAA dataset (Wang et al., 2024). Unlike prior few-shot synthesis that adapt model parameters using measured room data, we keep the trained model fixed and use $K = 8$ ordinary recordings per room only as inference-time conditioning.

As shown in Tab. 4, PHILAC achieves an EDC error of 2.797 dB, compared with 5.147 dB for FLAC and 6.375 dB for EigeNet. It obtains the lowest query-averaged error on five of the six reported metrics. The EDC improvement suggests that acoustic evidence extracted from unregistered recordings remains useful for predicting unseen paths in measured rooms, even without room-specific parameter updates. EigeNet achieves a slightly lower DRR error of 3.993 dB, compared with 4.041 dB for PHILAC. We report complementary results that weight each room equally in the Appendix D.8.

4.3 ABLATION STUDY

Main Components: Tab. 5 examines how the trained model uses blind audio, target geometry, and generative decoding. Removing the blind-audio features increases EDC error from 2.660 to

Table 5: Ablation of blind audio, target geometry, and stochastic decoding in PHILAC.

Intervention	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
w/o blind audio	9.346	10.968	65.02	6.602	0.5658	5.004
w/o visual/relative geometry	3.998	4.661	24.26	3.563	0.1597	3.172
w/ deterministic tokens	2.771	3.330	18.73	**1.857**	0.0930	**1.733**
Full model	**2.660**	**3.216**	**18.39**	2.031	0.0957	1.831

Table 6: Ablations of decoder memory, decay planner, and FiLM on unseen AcousticRooms scenes.

Memory	Planner	FiLM	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
–	–	–	$4.316_{\pm 2.978}$	$5.368_{\pm 3.773}$	$27.44_{\pm 23.26}$	$2.725_{\pm 2.097}$	$0.1619_{\pm 0.2000}$	$2.212_{\pm 1.721}$
–	✓	–	$6.805_{\pm 4.650}$	$8.192_{\pm 5.368}$	$48.26_{\pm 42.75}$	$4.605_{\pm 3.093}$	$0.2698_{\pm 0.2258}$	$3.616_{\pm 2.124}$
–	✓	✓	$2.789_{\pm 1.850}$	$3.415_{\pm 2.341}$	$18.08_{\pm 13.86}$	$2.026_{\pm 1.562}$	$0.1056_{\pm 0.0976}$	$1.727_{\pm 1.366}$
–	oracle	✓	$1.647_{\pm 0.714}$	$1.956_{\pm 0.852}$	$12.19_{\pm 8.13}$	$1.609_{\pm 1.148}$	$0.0684_{\pm 0.0568}$	$1.510_{\pm 1.161}$
✓	✓	✓	$2.660_{\pm 1.848}$	$3.216_{\pm 2.276}$	$18.39_{\pm 14.24}$	$2.031_{\pm 1.617}$	$0.0957_{\pm 0.0913}$	$1.831_{\pm 1.592}$

9.346 dB, showing that geometry alone provides insufficient information about the room response. Removing the panorama and query-relative geometry features increases the error to 3.998 dB, showing that acoustic evidence from the recordings must also be conditioned on the requested path.

Blind recordings do not uniquely determine the fine-scale structure of an unobserved RIR, which motivates the generative method. We thus assess it by replacing stochastic sampling with deterministic token prediction. The resulting EDC error increases from 2.660 to 2.771 dB, supporting the use of sampling for overall decay reconstruction.

Model Modules: We further examine the roles of each component in the model, as in Tab. 6. To isolate decay planning and FiLM, we retrain the corresponding variants without memory. In this setting, the model without a planner has an EDC error of 4.316 dB. Adding the planner alone increases the error to 6.805 dB, whereas passing its predicted decay to the decoder through FiLM reduces the error to 2.789 dB. This is a 35.4% improvement over the model without a planner. Late EDC error likewise decreases from 5.368 to 3.415 dB. The comparison shows that the decay plan improves synthesis when it is combined with FiLM.

Plan accuracy provides further room for improvement. Replacing the predicted decay with the query's true decay reduces EDC error from 2.789 to 1.647 dB and late EDC error from 3.415 to 1.956 dB. Although this oracle information is unavailable during inference, the result shows how strongly reconstruction depends on the quality of the plan. Taken together, these results show that decay planning is effective when delivered to the renderer through FiLM, and that more accurate plans could further improve RIR reconstruction.

5 CONCLUSION

Reconstructing a room's acoustic field from ordinary recordings is challenging because both the dry source signals and the recording locations are unknown. We addressed this problem by formulating blind acoustic field reconstruction as physics-anchored conditional completion, in which unlocalized recordings provide evidence of room-level reverberation and target geometry constrains the source–receiver path. Building on this formulation, PHILAC predicts a query-conditioned energy-decay profile and generates the target RIR through masked discrete diffusion of acoustic codec tokens. Experiments on unseen synthetic and real rooms show reductions in mean EDC error of 19.5% and 45.7%, respectively, over the strongest baselines, with no scene-specific adaptation on real rooms. Taken together, these results show that ordinary recordings can support acoustic field reconstruction at previously unobserved configurations, extending spatial audio rendering to settings where calibrated RIR measurements are unavailable.

USE OF LARGE LANGUAGE MODELS

During the experimental phase, large language models (LLMs) assist with code generation for method implementation, experimental setups, and data visualization under strict human guidance. During manuscript preparation, LLMs assist with grammar, sentence structure, stylistic refinement, and the identification and organization of related work. The human authors check the original sources and verify all literature summaries and citations. The methodology, experimental design, and scientific conclusions remain the responsibility of the human authors. All AI-assisted outputs undergo critical human review before inclusion in the manuscript.

ETHICS STATEMENT

This work studies the reconstruction of room impulse responses from acoustic recordings and scene geometry. It does not involve interventions with human participants or automated decisions about individuals. Because recordings used in practical applications may contain speech or other sensitive content, their collection, use, and release should respect consent, privacy, and applicable dataset licenses. The method could also be used to render convincing audio in a reconstructed space; such use should follow appropriate disclosure and consent practices.

REPRODUCIBILITY STATEMENT

The main text describes the problem formulation, model components, and evaluation protocol. The appendix provides implementation details, including model architecture, acoustic codec pretraining, training and inference settings, data processing, evaluation metrics, and baseline adaptations. We will release the code.

REFERENCES

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Proceedings of the twelfth language resources and evaluation conference, pp. 4218–4222, 2020.

Silvia Arellano, Chunghsin Yeh, Gautam Bhattacharya, and Daniel Arteaga. Room impulse response generation conditioned on acoustic parameters. In 2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 1–5. IEEE, 2025.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. Advances in neural information processing systems, 34:17981–17993, 2021.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. ICML, 2019.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. Soundstorm: Efficient parallel audio generation. arXiv preprint arXiv:2305.09636, 2023.

Amandine Brunetto. Few-shot acoustic synthesis with multimodal flow matching. arXiv preprint arXiv:2603.19176, 2026.

Amandine Brunetto, Sascha Hornauer, and Fabien Moutarde. Neraf: 3d scene infused neural radiance and acoustic fields. In International Conference on Learning Representations, volume 2025, pp. 92876–92899, 2025.

Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. Advances in Neural Information Processing Systems, 35:8896–8911, 2022.

Jiajian Chen, Jiakang Chen, Hang Chen, Qing Wang, Yu Gao, and Jun Du. Mean-rir: Multi-modal environment-aware network for robust room impulse response estimation. In 2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 1–7. IEEE, 2025.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. arXiv preprint arXiv:2210.13438, 2022.

James Eaton, Nikolay D Gaubitch, Alastair H Moore, and Patrick A Naylor. Estimation of room acoustic parameters: The ace challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(10):1681–1693, 2016.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100, 2020.

Diwei Huang, Kunyang Lin, Peihao Chen, Qing Du, and Mingkui Tan. Magic: Map-guided few-shot audio-visual acoustics modeling. arXiv preprint arXiv:2405.13860, 2024.

Chong Jing, Zitong Lan, Junan Zhang, and Zhizheng Wu. Eigenet: Geometry-informed multi-modal learning for few-shot novel view rir prediction. arXiv preprint arXiv:2605.28101, 2026.

Jean-Marc Jot and Antoine Chaigne. Digital delay networks for designing artificial reverberators. In Proceedings of the 90th AES Convention, number 3030, Paris, France, February 1991.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. Visage: Video-to-spatial audio generation. In International Conference on Learning Representations, volume 2025, pp. 46404–46424, 2025.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. Advances in neural information processing systems, 36:27980–27993, 2023.

Zitong Lan, Yiwei Tang, Yuhan Wang, Haowen Lai, Yiduo Hao, and Mingmin Zhao. Building audio-visual digital twins with smartphones. In Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services, pp. 825–838, 2026.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In International conference on machine learning, pp. 3744–3753. PMLR, 2019.

Kyung Yun Lee, Nils Meyer-Kahlen, Karolina Prawda, Vesa Välimäki, and Sebastian J Schlecht. Anyrir: Robust non-intrusive room impulse response estimation in the wild. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 15682–15686. IEEE, 2026.

Yuelei Li, Hyunjin Kim, Fangneng Zhan, Ri-Zhao Qiu, Mazeyu Ji, Xiaojun Shan, Xueyan Zou, Paul Liang, Hanspeter Pfister, and Xiaolong Wang. Visual acoustic fields. arXiv preprint arXiv:2503.24270, 2025.

Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. Advances in Neural Information Processing Systems, 36:37472–37490, 2023a.

Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. arXiv preprint arXiv:2309.15977, 2023b.

Jackie Lin, Jiaqi Su, Nishit Anand, Zeyu Jin, Minje Kim, and Paris Smaragdis. Gencho: Room impulse response generation from reverberant speech and text via diffusion transformers. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 22602–22606. IEEE, 2026.

Huadai Liu, Tianyi Luo, Kaicheng Luo, Qikai Jiang, Peiwen Sun, Jialei Wang, Rongjie Huang, Qian Chen, Wen Wang, Xiangtai Li, et al. Omniaudio: Generating spatial audio from 360-degree video. arXiv preprint arXiv:2504.14906, 2025a.

Xiulong Liu, Anurag Kumar, Paul Calamia, Sebastia V Amengual, Calvin Murdock, Ishwarya Ananthabhotla, Philip Robinson, Eli Shlizerman, Vamsi Krishna Ithapu, and Ruohan Gao. Hearing anywhere in any environment. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5732–5741. IEEE, 2025b.

Meyer-Kahlen Nils Lluís, Francesc. Blind spatial impulse response generation from separate room-and scene-specific information. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

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

Sheng Lyu, Yuemin Yu, and Chenshu Wu. Temporal modeling of room impulse response generation via multi-scale autoregressive learning. In INTERSPEECH, 2025.

Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. Advances in Neural Information Processing Systems, 35: 2522–2536, 2022.

Eloi Moliner, Jean-Marie Lemercier, Simon Welker, Timo Gerkmann, and Vesa Välimäki. Buddy: Single-channel blind unsupervised dereverberation with diffusion models. In 2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 120–124. IEEE, 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. Advances in Neural Information Processing Systems, 38:50608–50646, 2026.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5206–5210. IEEE, 2015.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

Karol J Piczak. Esc: Dataset for environmental sound classification. In Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015–1018, 2015.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The musdb18 corpus for music separation. 2017.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. Musdb18-hq-an uncompressed version of musdb18. (No Title), 2019.

Anton Ratnarajah, Ishwarya Ananthabhotla, Vamsi Krishna Ithapu, Pablo Hoffmann, Dinesh Manocha, and Paul Calamia. Towards improved room impulse response estimation for speech recognition. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, 2023.

Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. Av-rir: Audio-visual room impulse response estimation. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 27154–27165. IEEE, 2024.

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. Advances in Neural Information Processing Systems, 37:130136–130184, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. arXiv preprint arXiv:2508.10104, 2025.

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

Guaneen Tong, Johnathan Chi-Ho Leung, Xi Peng, Haosheng Shi, Liujie Zheng, Shengze Wang, Arryn Carlos O'Brien, Ashley Paula-Ann Neall, Grace Fei, Martim Gaspar, et al. Multimodal neural acoustic fields for immersive mixed reality. IEEE Transactions on Visualization and Computer Graphics, 31(5):3397–3407, 2025.

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

Chunxi Wang, Maoshen Jia, and Wenyu Jin. Daras: Dynamic audio-room acoustic synthesis for blind room impulse response estimation. IEEE Transactions on Audio, Speech and Language Processing, 2025a.

Lijun Wang, Yixian Lu, Ziyan Gao, Kai Li, Jianqiang Huang, Yuntao Kong, and Shogo Okada. Berp: A blind estimator of room parameters for single-channel noisy speech signals. IEEE Transactions on Audio, Speech and Language Processing, 33:2215–2231, 2025b.

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

Siyi Xie, Hanxin Zhu, Xinyi Chen, Tianyu He, Xin Li, and Zhibo Chen. Sonic4d: Spatial audio generation for immersive 4d scene exploration. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 11087–11095, 2026.

Shuyang Xu, Zhiyang Dou, Mingyi Shi, Liang Pan, Leo Ho, Jingbo Wang, Yuan Liu, Cheng Lin, Yuexin Ma, Wenping Wang, et al. Mospa: Human motion generation driven by spatial audio. Advances in Neural Information Processing Systems, 38:108786–108808, 2026.

Masaki Yoshida, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Scene2sound: Auditory-grounded soundscape generation for 3d gaussian worlds. arXiv preprint arXiv:2608.00463, 2026.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Masked audio generation using a single non-autoregressive transformer. In International Conference on Learning Representations, volume 2024, pp. 43249–43268, 2024.

APPENDIX

A ADDITIONAL METHOD DETAILS

A.1 AUDIO ENCODER

We center and RMS-normalize each recording before extracting its time–frequency features:

$$\widetilde{y}_k = \frac{y_k - \mu(y_k)}{\sigma(y_k) + \varepsilon}, \qquad X_k = F_{\mathrm{aud}}\big(\log\big(1 + |\,\mathrm{STFT}(\widetilde{y}_k)|\big)\big)\,. \tag{13}$$

The shared audio encoder F_{aud} combines self-attention, local convolution, and temporal pooling. Normalization reduces variation in recording level. Training with different source contents, equalization, and noise encourages the encoder to retain room cues despite changes in the sounds that excite the room. No dry-source estimate or context-source coordinate is required.

A.2 TEMPORAL ALIGNMENT AND DECODING

Let $n_q = \mathrm{round}(f_s \tau_q)$ be the sample index corresponding to the geometric delay, n_a a fixed canonical anchor, and S_Δ a zero-padded shift. Before encoding a training target, we align it to this anchor:

$$h_q^{\mathrm{can}} = S_{n_a - n_q}(h_q). \tag{14}$$

Given predicted tokens and gain, decoding restores the canonical waveform and then its target delay:

$$\widehat{h}_q = S_{n_q - n_a}[\widehat{g}_q\, \Pi_{n_a} D_\phi(Z_q)]\,, \tag{15}$$

where Π_{n_a} zeros samples before the canonical anchor.

A.3 LEARNING OBJECTIVES

For a corrupted token grid Z_t, let M_b denote its masked positions in codebook b. The masked-token objective is

$$\mathcal{L}_{\mathrm{mask}} = \mathbb{E}_{t,M}\left[\sum_{b\in\mathcal{B}} w_b \sum_{j\in M_b} \mathrm{CE}\Big(p_\theta(z_0^j \mid Z_t, C_q, d_q), z_0^j\Big)\right]. \tag{16}$$

We also predict the coarse codebook Z_q^c from a fully masked grid:

$$\mathcal{L}_{\mathrm{full}} = \mathrm{CE}\big(p_\theta(Z_q^c \mid Z_t = \texttt{[MASK]}, C_q, d_q), Z_q^c\big)\,. \tag{17}$$

The complete conditional objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\mathrm{mask}} + \lambda_g \mathcal{L}_{\mathrm{gain}} + \lambda_r \mathcal{L}_{\mathrm{room}} \\ + \lambda_e \mathcal{L}_{\mathrm{EDC}} + \lambda_t \mathcal{L}_T + \lambda_f \mathcal{L}_{\mathrm{full}}.\end{aligned} \tag{18}$$

B IMPLEMENTATION DETAILS

B.1 IMPLEMENTATIONS OF THE MODELS

Condition encoders. The recording encoder follows BERP's acoustic-feature and room-aggregation design. A 512-point Hann-window STFT with hop 256 yields log-one-plus magnitude features. A linear projection maps the 257 frequency bins to width 256. Four shared Conformer-style blocks (Gulati et al., 2020) combine eight-head self-attention, feed-forward layers, and a gated depthwise convolution of kernel size 15. Temporal pooling gives 16 tokens per recording. Sixteen learned queries then attend across the unordered recordings. The resulting 16 room tokens have no recording-order or context-position embeddings. Their mean feeds a probabilistic room-parameter head.

The geometry encoder receives a three-channel panorama of surface points expressed relative to the target source. It has four stride-two 5×5 convolutions with widths 32, 64, 128, and 256. Each convolution is followed by group normalization and SiLU. Adaptive 4×8 pooling produces 32 spatial tokens. An MLP embeds the target coordinate, source–receiver distance, and delay $\|q\|/343$ as one further token.

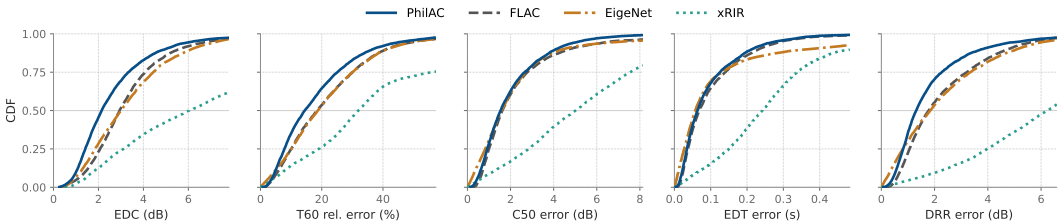


Figure 4: Per-query error distributions on unseen rooms.

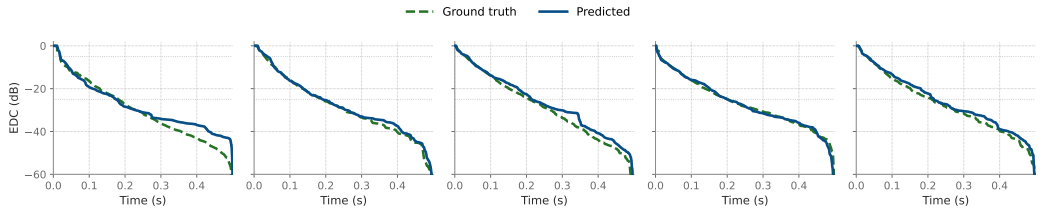


Figure 5: Illustrations of energy-decay curves.

Decay planner and renderer. The planner concatenates the pooled room and geometry features. A width-256 MLP predicts 63 positive increments. Their cumulative sum defines a 64-bin monotone EDC over 0.5 s. A differentiable fit gives T_{60}. We supervise the EDC with smooth L_1 loss above -60 dB and the logarithm of T_{60} with smooth L_1 loss. Their weights are 0.1 and 1.0. Eight pooled EDC controls and the fitted T_{60} form a nine-dimensional signal. The signal modulates every renderer block through FiLM. It is detached before entering the renderer. Room and geometry tokens form the cross-attention memory.

The renderer is a bidirectional Transformer decoder with 12 layers, width 768, 12 heads, and feed-forward expansion four. It sums frame-position and codebook-specific token embeddings. Separate categorical heads predict the eight codebooks. Another head predicts the gain token.

B.2 Training Details

Codec pretraining. The codec separates peak gain from waveform shape before encoding. Its causal waveform encoder has widths $(64, 96, 160, 256, 320)$ and two residual blocks per level. Four anti-aliased stride-two layers reduce a 0.5-s RIR to 500 frames of width 256. An exponential-moving-average residual quantizer uses eight codebooks with 512 entries each and decay 0.99. A separate 256-bin quantizer represents log peak gain over $[-8, 3]$. The causal decoder upsamples four times, restores gain, and reverses the geometric shift.

We first train the continuous autoencoder without quantization. We then warm-start the quantized RVQ stage. The committed base configuration uses AdamW (Loshchilov & Hutter, 2017) with learning rate 2×10^{-4} and weight decay 10^{-4}. It specifies batch size 32, 12 epochs, 2,000 warmup steps, cosine decay to 0.02 of the initial rate, bfloat16 mixed precision, and gradient clipping at 1. The complex spectral losses use FFT sizes 256, 512, and 1,024. The RVQ commitment weight warms up over 2,000 steps. We freeze the pretrained codec for conditional training.

Conditional training and sampling. The frozen codec supplies target RVQ indices and a gain class. We sample mask fractions with density $3t^2$, which emphasizes heavily masked inputs. Cross entropy is applied only at masked positions. Codebook weights $(4, 1, 1, 1, 1, 1, 1, 1)$ are normalized to mean one. Gain-token cross entropy has weight 0.25. The room-parameter negative log-likelihood has weight 0.1. We also supervise the planner as described above. In the final stage, a separate fully masked pass predicts the first codebook. Its loss has weight 0.5 and uses four examples per batch. The final conditional objective has no decoded-waveform loss.

We train the conditional generator in successive stages. We first learn masked token prediction from frozen codec targets with AdamW at 3×10^{-4} and batch size 32. We then fine-tune at 5×10^{-5}.

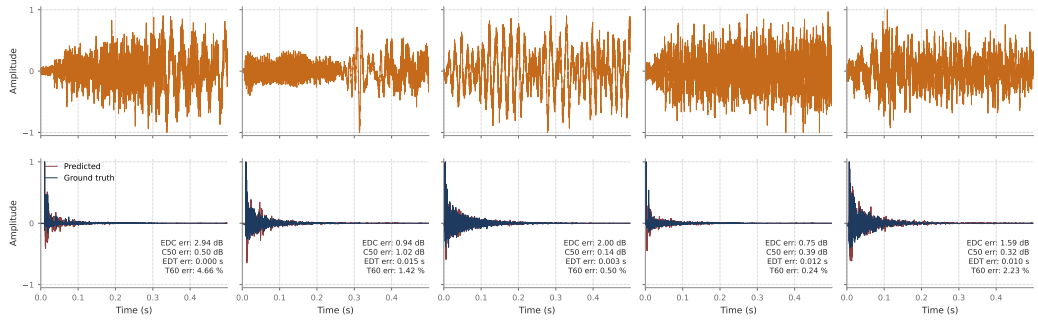


Figure 6: Illustrations of music observations and generated RIR waveforms.

This stage emphasizes high mask fractions and applies the codebook weights above. Next, we add the query-decay planner and block-wise FiLM. We train for six epochs at 5×10^{-5} for the existing parameters and 10^{-3} for the new modules. During this stage, the probability of supplying the true query EDC falls linearly from 1 to 0.2 over 20,255 updates. We finish with conditioning entirely on the predicted EDC. This last stage uses four NVIDIA H800 GPUs, batch size 16 per device, and AdamW at 2×10^{-5}. We use weight decay 10^{-4}, 200 warmup steps, cosine decay, bfloat16 mixed precision, and gradient clipping at 1.

Sampling starts with a fully masked 500×8 token grid. Each iteration predicts all masked positions in parallel. The least confident proposals are remasked separately within each codebook. We use two iterations and temperature 1.15. Gain is predicted greedily. The frozen codec decodes the token grid and restores the geometry-derived delay. The main result averages metrics across three fixed inference seeds.

B.3 Baseline Implementation and Adaptation

For the learned baselines adapted to our setting, we use the same room splits, target queries, and blind recordings as PHILAC. We replace measured context RIRs with these recordings and withhold context-source locations, while retaining the available target coordinates and depth geometry. We preserve each model's core architecture and retrain it on the common training split following its original learning objective and training recipe, with only the input changes required by our setting.

C Metrics

Let h and $\widehat{h}$ be the reference and predicted RIRs. We compute a Hann-window STFT with a 512-sample window and 128-sample hop. For $x \in \{h, \widehat{h}\}$, the frame energy and normalized backward-integrated decay are

$$e_x(j) = \sum_f |\mathrm{STFT}(x)_{f,j}|^2, \qquad D_x(j) = 10 \log_{10} \left(\frac{\sum_{u=j}^{J-1} e_x(u)}{\sum_{u=0}^{J-1} e_x(u)} \right). \tag{19}$$

Small numerical floors are applied inside the logarithm. This frame-energy EDC is different from backward integration of squared waveform samples. We evaluate its absolute difference on reference-selected frame sets

$$\mathcal{I} = \{j : D_h(j) > -60\}, \qquad \mathcal{I}_{\text{late}} = \{j : -60 < D_h(j) \le -10\},$$
$$\text{EDC error} = \frac{1}{|\mathcal{I}|} \sum_{j \in \mathcal{I}} |D_{\widehat{h}}(j) - D_h(j)|, \quad \text{Late EDC error} = \frac{1}{|\mathcal{I}_{\text{late}}|} \sum_{j \in \mathcal{I}_{\text{late}}} |D_{\widehat{h}}(j) - D_h(j)|. \tag{20}$$

Both errors are in dB. For $x \in \{h, \widehat{h}\}$, we fit a least-squares slope $a_x < 0$ to the points $(128j/f_s, D_x(j))$ for which $-25 \le D_x(j) \le -5$ dB. We extrapolate the fitted decay to 60 dB

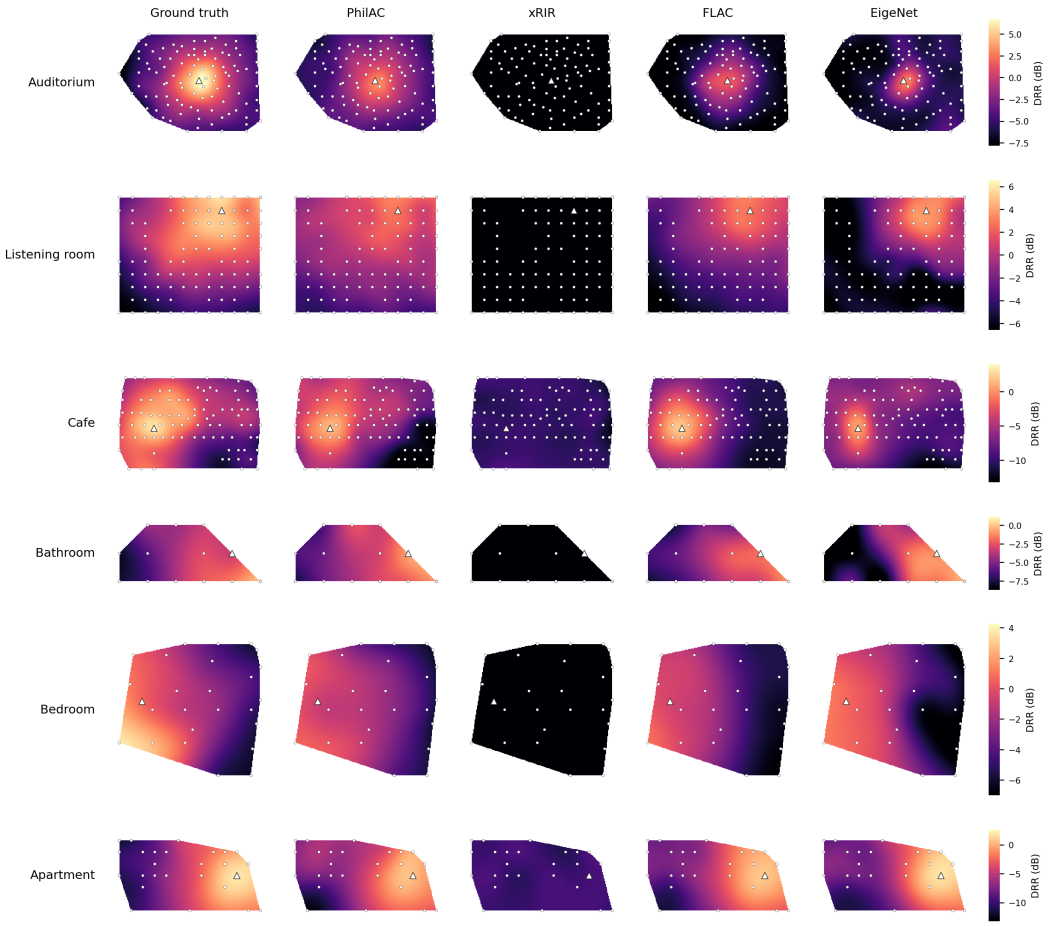


Figure 7: Spatial DRR fields among different rooms.

and report

$$T_{60}(x) = -\frac{60}{a_x}, \qquad \text{Relative } T_{60} \text{ error } (\%) = 100\frac{|T_{60}(\widehat{h}) - T_{60}(h)|}{\max(|T_{60}(h)|, 10^{-3})}. \tag{21}$$

The fit requires at least three valid frames. A missing fit yields an undefined error.

For C_{50}, the 50 ms split is measured from the first stored sample. Let $N = 8{,}000$ and $n_{50} = \lfloor 0.05 f_s \rfloor + 1$. We compute

$$C_{50}(x) = 10 \log_{10} \frac{\sum_{n=0}^{n_{50}-1} x[n]^2}{\sum_{n=n_{50}}^{N-1} x[n]^2}. \tag{22}$$

We compute EDT from the first -10 dB crossing of the waveform-energy decay,

$$W_x(n) = 10 \log_{10} \frac{\sum_{u=n}^{N-1} x[u]^2}{\sum_{u=0}^{N-1} x[u]^2}, \quad n_{-10}(x) = \min\{n : W_x(n) < -10\}, \quad \mathrm{EDT}(x) = 6\,\frac{n_{-10}(x)}{f_s}. \tag{23}$$

For DRR, $p_x = \arg\max_n |x[n]|$ anchors the direct window separately for each waveform. Let $\mathcal{J}_x$ contain samples from 1 ms before to 3 ms after p_x, with rounded endpoints and clipping to $[0, N-1]$. We compute

$$\mathrm{DRR}(x) = \mathrm{clip}_{[-60,60]}\left(10 \log_{10} \frac{\sum_{n\in\mathcal{J}_x} x[n]^2}{\sum_{n=0}^{N-1} x[n]^2 - \sum_{n\in\mathcal{J}_x} x[n]^2}\right) \text{ dB}. \tag{24}$$

If the direct or reverberant energy vanishes, the implementation assigns -60 or $+60$ dB, respectively. We report $|C_{50}(\widehat{h}) - C_{50}(h)|$, $|\mathrm{EDT}(\widehat{h}) - \mathrm{EDT}(h)|$, and $|\mathrm{DRR}(\widehat{h}) - \mathrm{DRR}(h)|$. Errors are computed per query, then averaged across queries. For stochastic methods, we first average per-query errors across draws. Undefined or non-finite values are omitted only from the corresponding metric's mean.

D Additional Experimental Results

D.1 Error Distributions

Figure 4 shows that PHILAC reaches lower EDC and relative T_{60} errors across a broad range of queries. The shift is also clear for DRR, while the C_{50} and EDT distributions are closer to those of FLAC and EigeNet. This pattern supports the mean improvements in Table 2 and shows that they extend across many target paths.

D.2 Qualitative Examples

Fig. 5 and Fig. 6 show some examples. The generated EDCs closely follow the reference curves through the early and middle decay, and capture the overall reverberant tail. Despite the different observed music signals, the generated RIRs reproduce the prominent early response and its subsequent energy decay. Fig. 5 displays waveform-energy EDCs.

Fig. 7 visualizes the DRR field in six unseen rooms. PHILAC recovers the reference's spatial structure, including high-DRR regions near the listeners and room-dependent gradients across the listening area. These coherent patterns show that blind recordings and query geometry support reconstruction of a spatially varying acoustic field.

D.3 Single Recording Synthesis Result

Table 7 shows that PHILAC leads all six metrics with a single blind recording. Its EDC error is 3.916 dB, compared with 4.544 dB for EigeNet, and its late EDC error is 4.692 versus 5.482 dB. Together with Table 3, this shows how additional unregistered recordings refine the inferred room evidence.

Table 7: Performance on unseen scenes with $K = 1$.

Method	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
Nearest	$6.059_{\pm 4.050}$	$7.487_{\pm 5.833}$	$41.99_{\pm 40.25}$	$3.986_{\pm 3.679}$	$0.2114_{\pm 0.2707}$	$9.726_{\pm 3.803}$
Blind T60+FDN (Jot & Chaigne, 1991)	$6.061_{\pm 4.048}$	$7.489_{\pm 5.828}$	$42.02_{\pm 40.36}$	$3.985_{\pm 3.679}$	$0.2116_{\pm 0.2706}$	$9.726_{\pm 3.803}$
FLAC (Brunetto, 2026)	$4.860_{\pm 4.268}$	$5.794_{\pm 4.758}$	$34.39_{\pm 38.26}$	$3.682_{\pm 3.821}$	$0.1897_{\pm 0.2207}$	$2.970_{\pm 2.223}$
EigeNet (Jing et al., 2026)	$4.544_{\pm 3.260}$	$5.482_{\pm 3.819}$	$31.84_{\pm 31.94}$	$3.434_{\pm 4.728}$	$0.1644_{\pm 0.1871}$	$2.875_{\pm 2.151}$
xRIR (Liu et al., 2025b)	$6.922_{\pm 4.536}$	$8.148_{\pm 5.306}$	$42.21_{\pm 35.27}$	$6.055_{\pm 5.850}$	$0.2796_{\pm 0.2113}$	$6.496_{\pm 3.433}$
Speech2RIR (Ratnarajah et al., 2023)	$23.120_{\pm 3.305}$	$31.921_{\pm 11.764}$	$85.87_{\pm 76.59}$	$5.905_{\pm 6.018}$	$0.4002_{\pm 0.2477}$	$3.643_{\pm 2.826}$
BUDDy (Moliner et al., 2024)	$13.621_{\pm 5.901}$	$14.647_{\pm 6.921}$	$42.72_{\pm 20.14}$	$15.639_{\pm 8.292}$	$0.5098_{\pm 0.3667}$	$14.558_{\pm 4.650}$
PHILAC	$\mathbf{3.916}_{\pm \mathbf{2.992}}$	$\mathbf{4.692}_{\pm \mathbf{3.528}}$	$\mathbf{27.55}_{\pm \mathbf{27.14}}$	$\mathbf{2.687}_{\pm \mathbf{2.040}}$	$\mathbf{0.1405}_{\pm \mathbf{0.1353}}$	$\mathbf{2.182}_{\pm \mathbf{1.788}}$

D.4 Different Dry Sources

Table 8 keeps the target queries and geometry fixed while changing the dry signals used to produce the blind recordings. On the three in-corpus source types, PHILAC achieves EDC errors of 2.660 dB for speech, 2.899 dB for broadband noise, and 3.129 dB for music. It leads on all six metrics in each setting. In particular, its broadband result remains close to its speech result, while FLAC and EigeNet show substantially larger errors.

The advantage extends to held-out sources. Common Voice speech yields an EDC error of 2.646 dB, nearly identical to the LibriSpeech result. The error rises to 3.874 dB for MTG-Jamendo music and 4.599 dB for ESC-50 sound events, reflecting the greater difficulty of these source shifts. Nevertheless, PHILAC retains the lowest error on every metric in all three held-out domains. These results

indicate that the inferred room representation transfers across recording corpora and remains useful even when the source type was absent from training.

Table 8: Performances on different source domains. Only the dry excitation changes.

Method	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
LibriSpeech test-other (in-corpus)						
FLAC (Brunetto, 2026)	3.305	4.055	21.98	2.508	0.1043	2.323
EigeNet (Jing et al., 2026)	3.375	4.152	21.92	2.721	0.1256	2.328
xRIR (Liu et al., 2025b)	6.922	8.149	42.22	6.056	0.2797	6.499
PHILAC	**2.660**	**3.216**	**18.39**	**2.031**	**0.0957**	**1.831**
broadband (in-corpus)						
FLAC (Brunetto, 2026)	12.751	14.759	71.85	11.195	0.9192	8.855
EigeNet (Jing et al., 2026)	9.914	11.472	68.83	8.101	0.5546	6.872
PHILAC	**2.899**	**3.565**	**18.71**	**2.090**	**0.1018**	**1.659**
MUSDB18-HQ test (in-corpus)						
FLAC (Brunetto, 2026)	5.962	7.001	43.32	4.600	0.2270	3.469
EigeNet (Jing et al., 2026)	6.560	7.685	48.14	5.066	0.2612	3.969
PHILAC	**3.129**	**3.804**	**22.68**	**2.149**	**0.1104**	**1.847**
Common Voice English (Ardila et al., 2020) (held-out speech)						
FLAC (Brunetto, 2026)	3.308	4.052	22.21	2.499	0.1038	2.300
EigeNet (Jing et al., 2026)	3.385	4.160	22.15	2.721	0.1252	2.365
PHILAC	**2.646**	**3.208**	**18.08**	**2.014**	**0.0965**	**1.802**
MTG-Jamendo test (Bogdanov et al., 2019) (held-out music)						
FLAC (Brunetto, 2026)	6.419	7.521	47.44	4.874	0.2572	3.645
EigeNet (Jing et al., 2026)	6.878	8.042	50.98	5.319	0.2831	4.132
PHILAC	**3.874**	**4.649**	**29.22**	**2.463**	**0.1274**	**1.987**
ESC-50 (Piczak, 2015) (held-out events)						
FLAC (Brunetto, 2026)	4.859	5.803	32.15	3.836	0.1821	3.019
EigeNet (Jing et al., 2026)	5.149	6.118	35.72	4.008	0.1981	3.445
PHILAC	**4.599**	**5.616**	**31.10**	**2.933**	**0.1529**	**2.203**

D.5 Ablations of Geometry encoder

Table 9 shows that the CNN reaches 2.660 dB EDC error, compared with 3.167 and 2.949 dB for frozen and fine-tuned DINOv3-S/16 (Siméoni et al., 2025). The query-relative depth panorama organizes wall boundaries and larger surface layouts spatially. Convolutional layers combine these features across scales, while query coordinates and direct-path timing provide explicit path cues. This makes the CNN well suited to the geometry used for acoustic prediction.

Table 9: Geometry encoder comparison.

Geometry encoder	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
DINOv3-S/16 frozen	3.167	3.768	19.54	3.012	0.1461	3.400
DINOv3-S/16 finetune	2.949	3.538	19.35	2.505	0.1231	2.818
CNN (ours)	**2.660**	**3.216**	**18.39**	**2.031**	**0.0957**	**1.831**

D.6 Ablations on Refinement Step

Table 10 compares refinement counts with the checkpoint and sampling temperature fixed. Two steps minimize EDC and late EDC error, at 2.662 and 3.217 dB. One step gives the lowest C_{50} and DRR errors, while four steps give the lowest EDT error. We use two steps for the main decay-oriented evaluation.

Table 10: Quality versus the number of refinement steps.

Steps	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
1	2.775	3.337	18.74	**1.860**	0.0934	**1.731**
2	**2.662**	**3.217**	**18.36**	2.034	0.0960	1.828
4	2.902	3.488	19.94	2.169	**0.0896**	1.886
8	3.523	4.202	21.71	2.575	0.0982	1.957
16	4.129	4.907	23.91	2.772	0.1032	1.850

D.7 Ablations on acoustic codec

Table 11 compares three frozen codec backends with separately trained conditional renderers. The RIR-specific codec achieves 2.660 dB EDC error, compared with 4.101 dB for DAC and 3.287 dB for EnCodec, and leads on the other five metrics. Direct-path canonicalization and RIR-specific temporal tokens provide an effective target for conditional completion.

Table 11: Ablations on different acoustic codec.

Token backend	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
DAC 16 kHz (Kumar et al., 2023)	$4.101_{\pm 2.046}$	$4.927_{\pm 2.486}$	$31.29_{\pm 18.53}$	$3.725_{\pm 3.248}$	$0.1727_{\pm 0.1269}$	$2.928_{\pm 1.601}$
EnCodec 24 kHz @ 6 kbps (Défossez et al., 2022)	$3.287_{\pm 1.874}$	$3.921_{\pm 2.243}$	$22.58_{\pm 33.40}$	$2.987_{\pm 3.058}$	$0.1312_{\pm 0.1174}$	$2.982_{\pm 1.748}$
PHILAC	$\mathbf{2.660}_{\pm 1.848}$	$\mathbf{3.216}_{\pm 2.276}$	$\mathbf{18.39}_{\pm 14.24}$	$\mathbf{2.031}_{\pm 1.617}$	$\mathbf{0.0957}_{\pm 0.0913}$	$\mathbf{1.831}_{\pm 1.592}$

D.8 Room-macro Zero-Shot Transfer

Table 12 complements the query-weighted HAA results in Table 4. Let $\mathcal{Q}_{r,m}$ contain the target queries in room r with a finite error $e_m(q)$ for metric m. We first average within each room and then give the four measured rooms equal weight:

$$\mu_{r,m} = \frac{1}{|\mathcal{Q}_{r,m}|} \sum_{q \in \mathcal{Q}_{r,m}} e_m(q), \qquad \mu_m^{\text{room}} = \frac{1}{4} \sum_{r=1}^{4} \mu_{r,m}, \qquad s_m^{\text{room}} = \sqrt{\frac{1}{3} \sum_{r=1}^{4} (\mu_{r,m} - \mu_m^{\text{room}})^2}. \tag{25}$$

The table reports $\mu_m^{\text{room}} \pm s_m^{\text{room}}$ after averaging stochastic draws within each query. PHILAC achieves 3.212 dB room-macro EDC error, compared with 5.392 dB for FLAC and 5.875 dB for EigeNet. It also leads on the other room-macro measures across the four measured rooms.

D.9 Spatial Behavior

Figure 8 groups per-query DRR, C_{50}, and EDT errors by source–receiver distance. Each violin shows the error distribution within a distance band. The comparison reveals how early-energy and reverberation estimates vary across short and long paths, complementing the dataset-wide means in Table 2.

D.10 Robustness to Observation Noise and Channel Effects

Table 13 varies the blind speech observations while keeping the model and target queries fixed. At the training SNR of 20 dB, PHILAC obtains 2.660 dB EDC error. The error rises to 4.029 dB at 10 dB

Table 12: Zero-shot transfer to four real HAA rooms with $K = 8$ (Room-macro results)

Method	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
FLAC (Brunetto, 2026)	$5.392_{\pm 2.310}$	$6.898_{\pm 3.705}$	$26.44_{\pm 18.43}$	$4.558_{\pm 1.493}$	$0.3442_{\pm 0.3637}$	$4.439_{\pm 0.449}$
EigeNet (Jing et al., 2026)	$5.875_{\pm 3.527}$	$7.764_{\pm 6.187}$	$29.81_{\pm 21.77}$	$4.910_{\pm 1.750}$	$0.3738_{\pm 0.4194}$	$4.277_{\pm 0.991}$
xRIR (Liu et al., 2025b)	$7.436_{\pm 4.962}$	$9.419_{\pm 6.151}$	$62.09_{\pm 79.67}$	$6.738_{\pm 5.418}$	$0.4201_{\pm 0.3656}$	$6.628_{\pm 3.645}$
PHILAC	$\mathbf{3.212}_{\pm 2.135}$	$\mathbf{3.583}_{\pm 2.194}$	$\mathbf{22.49}_{\pm 24.74}$	$\mathbf{3.880}_{\pm 0.694}$	$\mathbf{0.1734}_{\pm 0.0817}$	$\mathbf{4.135}_{\pm 0.531}$

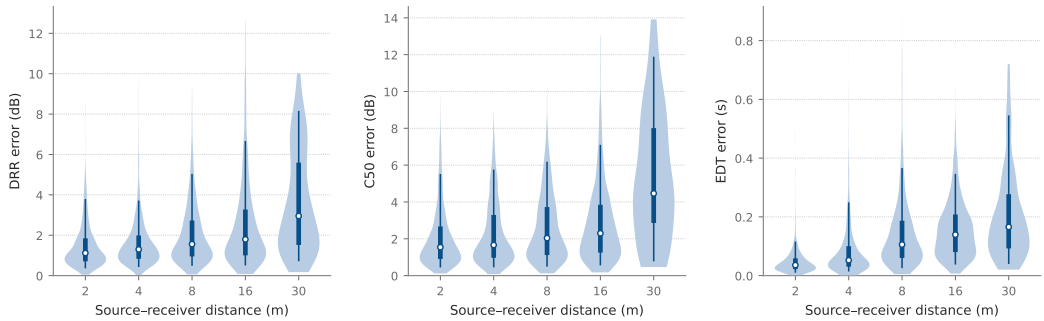


Figure 8: Acoustic errors versus source–receiver distance.

SNR and 7.740 dB at 0 dB. At 40 dB SNR, EDC error is 3.285 dB. These results show that PHILAC tolerates moderate observation noise, while reconstruction accuracy degrades substantially at very low SNR.

Eight-bit μ-law compression has almost no effect, with EDC error changing from 2.660 to 2.658 dB. Telephone-band filtering raises EDC error to 6.807 dB and increases the other errors as well. PHILAC therefore tolerates μ-law compression but degrades under severe bandwidth limitation.

Table 13: Noise and channel robustness of PHILAC.

Context observation	EDC ↓ (dB)	Late EDC ↓ (dB)	T_{60} rel. ↓ (%)	C_{50} err. ↓ (dB)	EDT err. ↓ (s)	DRR err. ↓ (dB)
Additive noise (speech)						
SNR 40 dB	$3.285_{\pm 2.256}$	$3.975_{\pm 2.661}$	$22.96_{\pm 19.31}$	$2.026_{\pm 1.551}$	$0.1050_{\pm 0.0977}$	$1.796_{\pm 1.504}$
SNR 20 dB	$2.660_{\pm 1.848}$	$3.216_{\pm 2.276}$	$18.39_{\pm 14.24}$	$2.031_{\pm 1.617}$	$0.0957_{\pm 0.0913}$	$1.831_{\pm 1.592}$
SNR 10 dB	$4.029_{\pm 2.596}$	$4.918_{\pm 3.187}$	$27.87_{\pm 16.02}$	$3.054_{\pm 2.233}$	$0.1251_{\pm 0.1338}$	$2.414_{\pm 2.019}$
SNR 0 dB	$7.740_{\pm 5.158}$	$9.270_{\pm 6.353}$	$38.76_{\pm 20.11}$	$6.044_{\pm 3.915}$	$0.2357_{\pm 0.2684}$	$3.173_{\pm 2.347}$
Channel (speech, SNR 20 dB)						
8-bit μ-law	$2.658_{\pm 1.847}$	$3.216_{\pm 2.278}$	$18.32_{\pm 14.22}$	$2.031_{\pm 1.613}$	$0.0955_{\pm 0.0915}$	$1.823_{\pm 1.589}$
Telephone 300–3400 Hz	$6.807_{\pm 4.011}$	$7.893_{\pm 4.488}$	$42.79_{\pm 34.33}$	$5.760_{\pm 2.567}$	$0.3232_{\pm 0.1543}$	$4.542_{\pm 2.220}$

D.11 INFERENCE COST

Table 14 measures steady-state batch-one inference on an NVIDIA RTX 4090. With two refinement steps, PHILAC takes 16.08 ms per query, or 62.18 serial queries per second. This is 16.3 times faster than 16-step FLAC and 1.5 times faster than EigeNet under the reported operating points. Parallel token updates produce a complete RIR in two renderer passes.

Table 14: Inference latency.

Method	Operating point	Mean ↓ (ms/query)	95% CI (ms)	P95 ↓ (ms)	Serial ↑ (q/s)	RTF ↓
FLAC	$N = 1$	261.60	[261.48, 261.73]	263.28	3.82	0.563
EigeNet	deterministic	24.28	[24.23, 24.33]	24.58	41.19	0.049
xRIR	32 GL iterations	53.62	[53.53, 53.72]	54.39	18.65	0.123
PHILAC	2 steps	**16.08**	[16.05, 16.13]	**16.29**	**62.18**	**0.032**

E DISCUSSIONS AND FUTURE WORK

Moving Sources and Listeners: PHILAC predicts RIRs at previously unobserved source positions relative to a fixed receiver. This setting supports practical applications such as spatial audio editing and video re-rendering from a fixed viewpoint, where existing recordings provide room evidence for placing new sounds. Our current formulation, however, does not cover scenarios in which both the source and the listener move during recording or rendering. Such motion changes the propagation paths over time and requires acoustic evidence to be integrated across receiver locations. Extending the framework to jointly varying source and receiver positions would broaden its applicability to mobile capture and interactive navigation.

RGB and Streaming Video: The current geometry branch uses a receiver-centered depth panorama to describe surrounding surfaces relative to each target source. Although this representation provides explicit geometric cues, it assumes that panoramic depth is available before synthesis. An important extension is to condition the model on ordinary RGB images, with scene structure inferred jointly from visual and acoustic observations. Video streams could further provide complementary views as the camera moves, allowing the scene representation to be updated as new evidence becomes available. Combining such incremental scene understanding with low-latency RIR generation could enable real-time spatial audio rendering from streaming video.

Scaling Evaluation to More Real Rooms: Our zero-shot evaluation on HAA rooms provides initial evidence that acoustic priors learned from synthetic data can transfer to real environments without room-specific parameter updates. Future work should explore larger measured-room datasets spanning a wider range of room sizes, layouts, furnishings, surface materials, and recording conditions. Such data would support more systematic evaluation of sim-to-real transfer and help identify which environmental factors most strongly affect blind acoustic field reconstruction. They could also enable training with more diverse real acoustic evidence.

Continuous Trajectories and Spatial Consistency: Our evaluation primarily measures reconstruction quality at individual query positions. Although queries share room evidence, their RIRs are generated separately, without an explicit constraint on consistency along a continuous trajectory. Accurate predictions at isolated positions therefore do not by themselves establish perceptually coherent rendering as a source moves. Future work could jointly generate neighboring queries or condition each prediction on preceding responses, encouraging coherent evolution of reflection structure, energy decay, and gain. Trajectory-level evaluation would further assess temporal continuity and audible artifacts during motion, complementing the current per-query acoustic metrics.

Generalization to Everyday Sound Sources: Our training sources consist of speech, music, and broadband noise, and the source-domain evaluation shows that performance still depends on the sounds used to obtain acoustic evidence. In particular, the EDC error increases to 4.599 dB on held-out ESC-50 sound events, compared with 2.660 dB for speech, 3.129 dB for music, and 2.899 dB for broadband noise (Table 8). Although PHILAC remains the strongest evaluated method on ESC-50, this gap indicates substantial room for improvement on everyday sounds. A useful next step is to broaden source diversity during training and investigate how event duration, spectral coverage, and temporal sparsity affect the room cues available to the encoder. Improving robustness to these variations would make acoustic field reconstruction more reliable when context recordings contain incidental environmental sounds.