Audio Engineering Society

Convention Paper

Presented at the AES 161st Convention
2026 October 30-November 1, Nashville, TN, USA

*This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.*

# Where Can Decay Losses Act? Characterizing EDR Supervision Across Training and Sampling in Diffusion RIR Synthesis

Vlad Baran[1,2], Jack Kelly[1], and Wieslaw Woszczyk[1]

[1]*McGill University*
[2]*CIRMMT*

Correspondence should be addressed to Vlad Baran (vlad.baran@mail.mcgill.ca)

**ABSTRACT**

Differentiable energy-decay losses have been shown to improve deterministic artificial reverberation algorithms, such as feedback delay networks. In this work, we investigate whether these benefits transfer to diffusion-based room impulse response (RIR) generation. A v-prediction diffusion U-Net operating on complex spectrograms is trained on approximately 10,000 measured RIRs and evaluated on 384 unseen room impulse responses from six receiver placements. The proposed Energy Decay Relief (EDR) criterion is incorporated in three stages of the generation pipeline: direct training, post-training fine-tuning, and inference-time guidance. The baseline diffusion model already reproduces the conventional room acoustic parameters ($RT_{60}$, $EDT$, and $C_{50}$) with errors below their perceptual thresholds, leaving the late reverberation tail as the primary remaining source of error. Although the EDR criterion successfully influences the targeted decay characteristics in all three regimes, it does not improve the overall quality of the generated responses. Direct training produces only marginal improvements while slightly increasing the reverberation time, fine-tuning improves the training objective by 1.1 dB without improving the generated responses, and inference-time guidance enhances the late reverberation tail only at the expense of perceptually noticeable reductions in clarity. These conclusions are supported by paired statistical comparisons over the full hold-out dataset.

## 1 Introduction

Room Impulse Responses (RIRs) measurements are representations that describe an environment's physical response to excitation over time. These representations are essential for various applications, including music production, virtual and augmented reality, immersive computer games, and acoustical characterization of rooms and concert halls [1, 2, 3]. A persistent issue noted in the literature [4, 5] is the "fidelity gap" between simulated and real-world RIRs. Measuring room impulse responses is a resource-intensive process that restricts the reproduction to the measured positions. Conversely, simulated room impulse responses can be generated efficiently for a wide range of environments but often fail to capture the complexity and acoustic realism of real-world spaces. Traditional RIR reconstruction methods are primarily based on mathematical formulations, including compressed sensing techniques and numerical solutions of the acoustic

wave equation [6, 7]. While effective in controlled settings, these approaches often face challenges such as mishandling directional source energy and excessive brightness when dealing with complex acoustic environments and highly reverberant spaces [8]. More recently, deep learning methods have been introduced to overcome these limitations, with architectures such as Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GAN) demonstrating improved reconstruction performance. For example, GAN-based approaches have been successfully applied to bandwidth extension in array processing tasks, while Physics-Informed Neural Networks (PINNs) incorporate acoustic constraints to enhance prediction accuracy [9, 10].

Diffusion Denoising Probabilistic Models (DDPMs) have also emerged as a promising framework for sound field reconstruction, leveraging probabilistic generation to produce accurate acoustic representations [11].Lately, generative modeling has been explored by[12, 13, 14, 15] for the synthesis of room acoustics and as a data augmentation strategy or position interpolation. This work explores the improvement of Denoising Diffusion Implicit Models (DDIM) using differentiable objective metrics [16]. To preserve the complete acoustic information contained in each measurement, the room impulse responses are represented in the complex short-time Fourier transform (STFT) domain as two-channel complex STFT representations (real and imaginary parts). A diffusion model is then trained to generate RIRs from these representations while incorporating the Energy Decay Relief (EDR) criterion. EDR is an objective measure that describes how acoustic energy decays over time across frequency bands, providing a perceptually meaningful description of reverberation [17]. Building on the work of Mezza et al. [18], the present study explores the use of the EDR term within a diffusion U-Net architecture to encourage the generation of high-fidelity room impulse responses.

### 1.1 Diffusion Networks

Diffusion models have recently emerged as a powerful approach for RIR generation by learning the statistical distribution of acoustic responses and progressively denoising random noise into realistic samples. Unlike deterministic neural networks, such as convolutional encoder-decoder architectures that learn a direct mapping from the parameters to a single room impulse response, diffusion models learn the underlying probability distribution of the training data and generate new RIRs through an iterative denoising process. Denoising Diffusion Probabilistic Models(DDPM) represent a class of generative models that learn to reverse a gradual, multi-step noising process to match a target data distribution [11]. This framework is built upon two distinct stages: a forward noising process and a reverse denoising process. In the forward process, a real data point $x_0$ is progressively corrupted by adding diagonal Gaussian noise at each timestep (t) according to a variance schedule until it approximates a simple prior distribution, such as isotropic Gaussian noise $x_T$. The reverse process is defined by a Markov chain parameterized by a neural network that learns to eliminate the added noise step-by-step, effectively reconstructing the original data from a latent variable. To overcome the slow sampling speeds inherent in standard Markovian DDPMs, researchers developed DDIMs, which utilize non-Markovian diffusion processes [16]. This generalization allows for deterministic generative trajectories that can produce high-quality samples in an order of magnitude fewer steps (e.g., 20–100 steps instead of 1,000) while sharing the same training objective as DDPMs. Additionally, specialized diffusion architectures like DiffWave have been proposed for parallel raw audio synthesis, matching the quality of strong autoregressive models like WaveNet while being significantly faster at inference [19, 20].

Regarding impulse response generation, diffusion networks have been applied to several related tasks, including interpolation between impulse responses[21], completion of partially observed impulse responses[22], and text-conditioned generation of room impulse responses[23]. These approaches demonstrate the potential of diffusion models for generating and reconstructing room impulse responses under different conditioning schemes. While the results vary across these studies none of the approaches present a controlled ablation of a decay objective across training regimes.

## 2 Motivation

In their paper [18], Mezza et al. discuss automatic parameter tuning of feedback delay-networks. In their approach the researchers focused on the problem that existing models often match global decay but fail to capture the frequency-specific behavior of real rooms.

The authors demonstrated that EDR provides an effective differentiable objective for optimizing deterministic artificial reverberation algorithms. Building on this work, the present study investigates whether the same motivated criterion can be successfully transferred to diffusion-based room impulse response generation. EDR is a time-frequency distribution that generalizes Schroeder's Energy Decay Curve (EDC) to multiple frequency bands [17]. It represents the total amount of signal energy remaining in a room impulse response at a given time and frequency band. It is calculated using the following formula:

$$\mathrm{EDR}(t_n, f_k) = \sum_{m=n}^{M} |H(m,k)|^2 \tag{1}$$

where $H(m,k)$ denotes the $k$-th frequency bin of the short-time Fourier transform (STFT) at time frame $m$, and $M$ is the total number of STFT frames. The quantity $\mathrm{EDR}(t_n, f_k)$ represents the total remaining energy in the impulse response at time $t_n$ within the frequency band centered at $f_k$.

## 3 Proposed Method

Both networks used in this study employ the same denoising architecture and operate on a Short-Time Fourier Transform representation of each RIR. Before being processed by the network, each RIR is converted into a standardized format by peak-normalizing the audio, aligning the direct sound to a common starting point, and computing a 2048-point STFT with a hop size of 256 samples. The resulting complex spectrogram is represented by its real and imaginary components and contains the complete decay of the impulse responses.

The denoiser is based on a U-shaped convolutional neural network Figure. 1, an architecture that has been widely adopted in image generation and more recently in diffusion models for audio. It was originally introduced for biomedical image segmentation, the U-Net is an efficient encoder-decoder framework characterized by a contracting path to capture context and a symmetric expansive path for precise localization [24]. The "U" shape arises from skip connections that concatenate high-resolution feature maps from the encoder directly to the decoder stages. These connections are critical as they allow the network to retain fine-grained spatial or temporal information that might otherwise be lost during the downsampling process. For audio applications, architectures such as the Deep Complex U-Net (DCUNet) have been adapted to operate directly on complex-valued Short-Time Fourier Transform (STFT) representations [25].

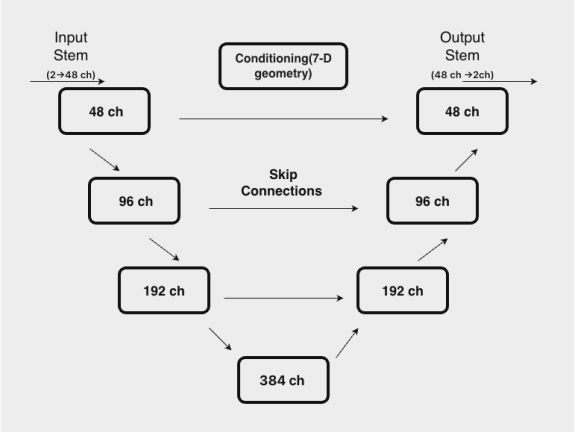



**Fig. 1:** U-Net architecture with 7-D conditioning vector.

This architecture is particularly well suited to room impulse responses because they contain information at several different time scales. The beginning of an RIR consists of the direct sound and a series of early reflections, which require the network to preserve fine time-frequency details. In contrast, the later part of the signal is dominated by the low energy sound that decays gradually. The encoder progressively reduces the resolution of the input, allowing the network to capture this larger context, while the decoder reconstructs the full-resolution spectrogram. Skip connections between the encoder and decoder ensure that fine details from the early reflections are preserved throughout the reconstruction process. Each convolutional block in Fig,1 uses residual connections together with Group Normalization and SiLU activation functions to improve training stability. The network is conditioned on the diffusion timestep and a seven-dimensional spatial vector (receiver position, microphone orientation, and source-to-receiver distance) while a time-axis positional encoding preserves the temporal structure of the room impulse response.

The resulting feature representation is then passed to the decoder, where the original spectrogram resolution is progressively reconstructed. A complex STFT representation was chosen because it preserves both the

magnitude and phase of the signal, providing a complete description of the room impulse response. In contrast, magnitude-only representations discard phase information, making it necessary to estimate or reconstruct the missing phase during synthesis [26]. By predicting the real and imaginary components of the STFT directly, the network jointly models magnitude and phase, allowing the room impulse response to be recovered with a straightforward inverse STFT. This approach improves waveform fidelity while avoiding the need for iterative phase reconstruction algorithms or separate neural vocoders, which need to be trained separately [27]. In this architecture, we adopt the v-prediction parameterization proposed by [28] together with the Min-SNR loss weighting strategy introduced by [29]. Rather than predicting the noise added to the signal or the clean signal, $v$-prediction trains the network to estimate a linear combination of the two. The $v$-parameterization used in this work is defined as

$$v_t = \sqrt{\bar{\alpha}_t}\varepsilon - \sqrt{1-\bar{\alpha}_t}x_0,$$

where $v_t$ is the velocity prediction target at diffusion timestep $t$, $x_0$ represents the original clean room impulse response, $\varepsilon$ is the Gaussian noise added during the forward diffusion process, and $\bar{\alpha}_t$ represents the cumulative noise schedule that determines the relative contribution of the signal and noise components at timestep $t$. Min-SNR loss weighting formulation prevents the learning objective from becoming dominated by either highly noisy timesteps, where the clean signal information is weak, or nearly clean timesteps, where the noise prediction becomes trivial.

### 3.1 Dataset

The dataset used in this work consists of approximately 10,000 RIRs measured in the Redpath Hall at McGill University using two 64-capsule Eigenmike EM64 spherical microphone arrays with identical orientations. The room impulse response dataset was acquired in Redpath Hall at McGill University, a performance corresponding to a volume of approximately 4825 $m^3$. During the measurements the hall had a reverberation time of approximately $RT_{60} = 2.2$ s. An open staircase connecting the main hall to the basement and upper loft acoustically coupled these spaces, contributing to the measured reverberation characteristics. Measurements were collected over a grid of 162 unique receiver positions, defined by nine depths, three lateral spacings,



**Fig. 2:** Redpath Hall, McGill University $RT_{60} \approx 2.2$ s at the measured positions.

**Table 1:** Hold Out positions (64 channels each).

| Position | Height (m) | Source (m) | East wall (m) |
|---|---|---|---|
| EM64L_P2_2 | 1.5 | 8.21 | 8.8 |
| EM64LH_P2_2 | 2.0 | 8.19 | 8.8 |
| EM64LHH_P2_2 | 2.5 | 8.21 | 8.8 |
| EM64R_P2_2 | 1.5 | 8.21 | 3.8 |
| EM64RH_P2_2 | 2.0 | 8.19 | 3.8 |
| EM64RHH_P2_2 | 2.5 | 8.21 | 3.8 |

two sides of the hall, and three measurement heights, providing a diverse sampling of the room's acoustic field. During preprocessing, each 64-channel recording is separated into individual mono impulse responses, trimmed, RMS-normalized, and automatically labeled according to its measurement position and capsule number. During training 6 positions are excluded from the training data. This totals 384 unseen validation samples that are also used in the statistical analysis. In Table 1 the measurements of 6 positions are shared (Left, Right, 3 height settings, 64 capsules/position)

### 3.2 Statistical analysis

To provide a statistically meaningful evaluation, all models are tested on the same held-out measurement grid. Each model generates one RIR per receiver using

identical conditioning and diffusion seeds, and every generated response is compared with its corresponding measured reference. Model performance is evaluated across the entire dataset rather than from individual examples, allowing the overall error distribution to be characterized. Following [30], perceptual thresholds are taken as approximately 5% for $RT_{60}$ and $EDT$, and 1 dB for $C_{50}$. For the late-tail level ($EDC@2$ s), no standardized threshold exists; therefore, a value of approximately 1 dB is used as a heuristic reference. Throughout this work, differences are considered meaningful only when they are both statistically significant and exceed the corresponding perceptual threshold. Pairwise comparisons between models are performed using the paired Wilcoxon signed-rank test, which determines whether one model consistently produces smaller errors than another without assuming normally distributed data. In addition to reporting the median error and the fraction of improved positions, the analysis distinguishes systematic bias, which may be addressed through improved training objectives, from the natural variability observed across different measurement positions.

## 4 Evaluation

To establish a baseline, the distribution of room acoustic metrics produced by the standard diffusion model is first compared with the measured room impulse responses. The baseline's systematic biases are median errors of $+1.4\%$ for RT60 (JND $\approx 5\%$), -0.5 dB for C50 (JND $\approx 1$ dB), and +0.09 s for EDT.

The most relevant bias is the late-tail level, is at EDC at 2s, where the measured references have decayed to approximately -54 dB ($\Delta$EDC@ 2s median +1.1 dB). The EDR criterion is evaluated in three application regimes: direct training, post-training fine-tuning, and inference-time guidance. Direct training compares a baseline v-prediction model against an identical model whose objective includes an additional EDR loss. Along all experiments the two networks differ only in the loss formulation. In the baseline $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon$ is the compressed spectrogram $x_0$ corrupted with Gaussian noise $\varepsilon$ at diffusion step $t$, and $v_t = \sqrt{\bar{\alpha}_t}\, \varepsilon - \sqrt{1-\bar{\alpha}_t}\, x_0$ is the velocity target, a signed combination of signal and noise that remains well-conditioned across the full noise range. The network prediction $\hat{v}_\theta(x_t, t)$ is compared to this target through the mean-squared error $\|\cdot\|_2^2$, taken over all time-frequency cells of the two-channel (real and imaginary) spectrogram. The weight $w_t = \min(\mathrm{SNR}_t, \gamma)/\mathrm{SNR}_t$ is the Min-SNR schedule, which caps the training effort spent on low-noise steps so that no region of the noise range dominates optimization; we use $\gamma = 5$. As such the baseline model's loss becomes:

$$\mathscr{L}_{\mathrm{v-pred}} = w_t \left\| v_t - \hat{v}_\theta(x_t, t) \right\|_2^2 \tag{2}$$

While the loss for v-pred+EDR is:

$$\mathscr{L}_{\mathrm{v-pred+EDR}} = w_t \left\| v_t - \hat{v}_\theta(x_t, t) \right\|_2^2 + \lambda\, \bar{\alpha}_t\, \mathscr{L}_{\mathrm{EDR}} \tag{3}$$

The EDR-supervised model extends this objective with a decay term. $\mathscr{L}_{\mathrm{EDR}}$ is the normalized-Huber EDR distance between the energy-decay surface of $\hat{x}_0$ and that of the measured reference $x_0$. The factor $\bar{\alpha}_t$ weights each sample's contribution by its noise level: it approaches 1 at low noise and vanishes at high noise. The constant $\lambda$ sets the overall strength of the decay term; it is fixed ($\lambda = 4 \times 10^{-4}$) and calibrated so that the EDR gradient contributes approximately 5–10% of the total gradient norm.

### 4.1 Direct Training

The two models are trained in parallel using the same parameters differing only in the objective: the baseline (v-pred) uses the v-prediction loss alone, while the EDR-supervised model (v-pred + EDR) adds the decay term of Eq. 3, weighted by $\lambda$. This experiment tests whether the EDR loss term applied throughout training yields generated impulse responses with more accurate decay than v-prediction training alone. Both models generate one RIR per held-out positions receiver (384 positions) under matched conditioning and noise seed. Statistically the per-position errors against the matched measured references are compared with Wilcoxon tests.

### 4.2 Post-Training Fine-Tuning

Both models start from the same converged v-pred checkpoint and continue optimization for 20 epochs at a low constant learning rate($10^{-5}$) with identical schedules and data. The treatment adds the $\bar{\alpha}_t$-weighted EDR term with $\lambda$ recalibrated to the same gradient-share band. The control continues with the v-loss alone isolating the effect of the EDR loss from that of the additional training itself. This regime tests whether the criterion can refine an already-formed model.

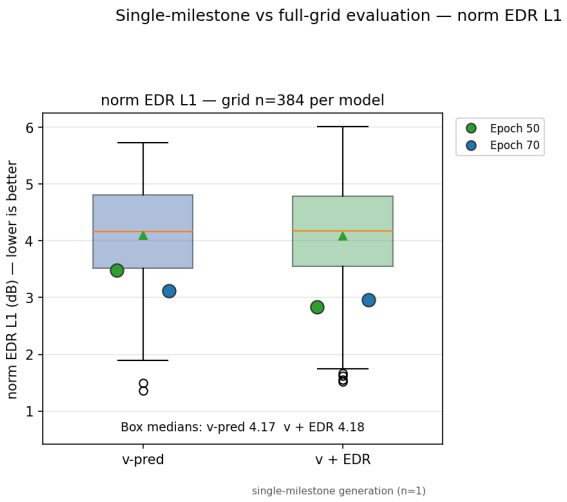



**Fig. 3:** Single generations (dots; one conditioning vector, one seed) suggest the EDR loss reduces decay-shape error by 0.4 dB; the full-grid distributions (boxes; n=384) show the models are indistinguishable (medians 4.17 vs 4.18, Wilcoxon p=0.57).

### 4.3 Inference-Time Guidance

For inference-time guidance, the trained diffusion model remains unchanged and the EDR criterion is applied only during the sampling process. This experiment is inspired by the work developed in [31, 32]. At each denoising step, the generated RIR is gently adjusted toward the target decay characteristics by modifying the sampling trajectory. The strength of this correction is controlled by a single parameter $s$, where small values produce weak adjustments and larger values impose a stronger influence from the EDR criterion. The value of $s$ therefore controls the trade-off between maintaining the original characteristics learned by the diffusion model and enforcing the desired decay behavior, without requiring additional model training. The strength s is selected by a sweep over 64 held-out positions using norm EDR L1 and $\Delta EDC@2s$.

## 5 Results

The first result of this study, Fig 3. demonstrates the importance of evaluating generative room impulse response models using a statistically meaningful protocol rather than individual examples. When the two models are compared using a single generated room impulse

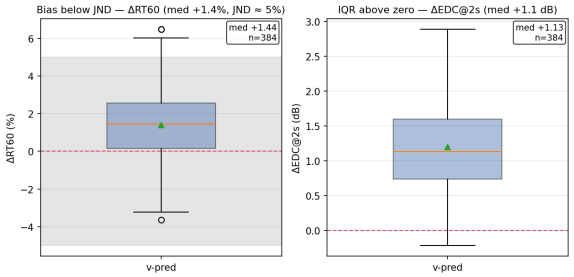



**Fig. 4:** $v$-prediction baseline evaluated against matched references ($n$=384). Dashed line: perfect agreement. Signed median errors: $\Delta RT60 = +1.44\%$ (JND $\approx$5%), $\Delta EDT = +0.09$ s, $\Delta C50 = -0.49$ dB (JND $\approx$1 dB), $\Delta$EDC@2s$= +1.13$ dB; median decay-shape error 4.17 dB (norm EDR $L_1$). Conventional decay metrics remain below their perceptual thresholds, while the late-tail level exhibits the only significant systematic bias.

response, the model trained with the EDR loss appears to produce a noticeable improvement over the baseline. However, this conclusion changes completely when the evaluation is extended to the full set of 384 unseen validation samples.

Across the complete dataset the performance of the two models is statistically indistinguishable. This finding highlights that room impulse response generation exhibits considerable variability between receiver positions, making individual examples unreliable indicators of overall performance. Consequently, all subsequent analyses in this work are based on paired statistical comparisons over the entire holdout set.

Figure 4. establishes the performance of the baseline diffusion model before introducing any energy-decay supervision. By comparing the generated and measured room impulse responses across all holdout samples, it becomes clear that the model biases remain below their perceptual thresholds. The only region of the RIR we found to be problematic is the EDC at 2 seconds which has an inter-quartile range entirely above zero.

Figures 5 and 6 present the result of the fine-tuning experiment. Figure 6. shows the normalized EDR error of the network's denoising estimates on held-out data throughout fine-tuning. The EDR-supervised model improves from 3.35 to 2.21 dB, while the control model remains essentially unchanged, demonstrating

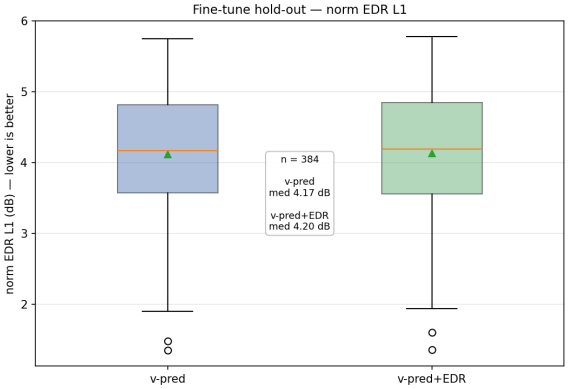



**Fig. 5:** After 20 epochs of fine-tuning. The median normalized EDR error remains essentially unchanged (4.17 vs. 4.20 dB)

that the EDR objective is successfully optimized during training. Figure 5 same metric on the generated room impulse responses over the full hold-out set. Despite the improvement observed during training, the generated outputs remain statistically unchanged (median norm EDR $L_1$: 4.17 vs. 4.20 dB), indicating that the optimization does not transfer to the sampled responses. This suggests that improving the intermediate denoising estimates does not necessarily improve the final room impulse responses.

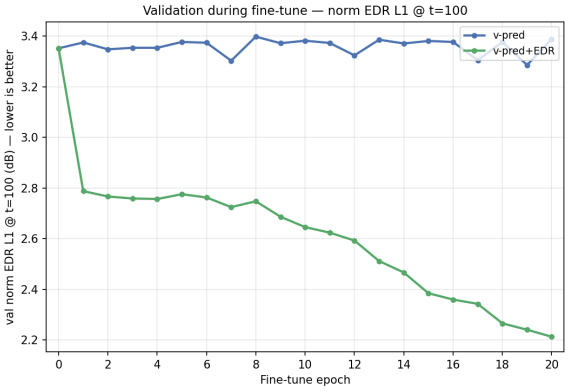



**Fig. 6:** Validation norm EDR $L_1$ measured at a fixed diffusion timestep ($t = 100$) during the 20-epoch fine-tuning. The baseline remains nearly constant ($\approx$3.35 $dB$), whereas the EDR-supervised model improves to 2.21 dB.

Figure 7. represents inference-time guidance, where the EDR criterion is applied during the diffusion sampling process rather than during training. Increasing the guidance strength consistently improves the accuracy of the late reverberation tail, but this improvement is accompanied by a progressive deterioration of early reflection clarity.

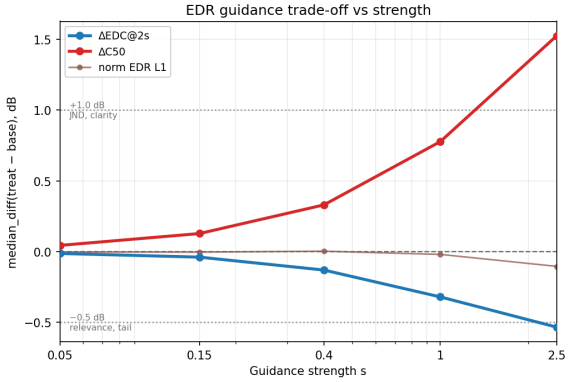



**Fig. 7:** EDR guidance controlled by the **s** term, late reverberation improves with the cost of clarity, no point yields a net improvement.

**Table 2:** EDR-criterion outcomes across the three application regimes. Cells: median change in absolute error relative to each regime's matched control (negative = closer to the measured reference). Paired Wilcoxon signed-rank tests; $n$=384 positions for Direct and FT, $n$=64 for guidance; † = not significant ($p$>0.05), all other cells $p<10^{-4}$. Controls: v-prediction baseline (Direct), v-only fine-tune (FT), unguided sampling (Guid., reported at $s$=2.5).

| | Direct | FT | Guid. $s$=2.5 |
|---|---|---|---|
| ΔEDC@2s (dB) | −0.17 [75%] | +0.19 [1%] | −0.53 [92%] |
| EDR $L_1$ (dB) | −0.03 [52%]† | +0.04 [31%] | −0.10 [63%]† |
| \|ΔRT60\| (%) | +0.79 [24%] | +0.62 [20%] | −1.18 [72%] |
| \|ΔC50\| (dB) | −0.02 [52%]† | +0.03 [33%] | +1.53 [9%] |
| Verdict | marginal trade | no transfer | cost > gain |

Table 2 summarizes the performance of the three EDR regimes: direct training, post-training fine-tuning, and inference-time guidance. Direct training produced only a marginal reduction in late-tail error (ΔEDC@2s = −0.17 dB at 75% of positions), but this came with a slight increase in RT60 error. Post-training fine-tuning failed to transfer its training improvements to the generated RIRs, slightly degrading both the tail level and

decay-shape metrics. Inference-time guidance achieved the largest improvement in tail accuracy ($-0.53$ dB at 92% of positions); however, this was accompanied by a substantial increase in clarity error ($+1.53$ dB). Overall, the EDR loss implemented consistently influenced the decay characteristics in every experiment, but none of the three application strategies produced a net improvement in the realism of the generated room impulse responses.

## 6 Conclusion

We investigated whether the Energy Decay Relief (EDR) criterion, previously shown to improve deterministic artificial reverberation algorithms, can also improve diffusion-based room impulse response generation. The EDR loss was incorporated in three stages of the generation pipeline: direct training, post-training fine-tuning, and inference-time guidance. Each approach was evaluated using paired statistical comparisons over 384 unseen room impulse responses.

The baseline diffusion model already reproduced the conventional room acoustic parameters with errors below the perceptual thresholds established by [30], leaving the late reverberation tail as the primary remaining source of error. The fine-tuning experiment revealed the main limitation of the EDR criterion. Although the network learned the EDR objective and reduced the training-time decay error of its intermediate denoising estimates by approximately 1.1 dB, this improvement did not translate to the final generated room impulse responses. This suggests that optimizing intermediate denoising estimates is not sufficient to improve the complete diffusion sampling process that ultimately determines the generated response. Applying the EDR criterion during inference improved the late reverberation tail for most positions. However, these improvements came at the expense of reduced clarity, with noticeable objective degradation occurring before the tail improvements became meaningful.

Taken together, these results provide a systematic evaluation of the EDR criterion within a diffusion-based room impulse response generator. Although the criterion successfully modifies the decay behavior of the generated responses, none of the investigated application strategies produced a consistent improvement in overall realism. These findings demonstrate that the success of EDR in deterministic artificial reverberation does not directly transfer to diffusion models and help define the practical limitations of applying perceptually motivated decay objectives to iterative generative frameworks. Finally, evaluating performance over a large set of matched measurements proved essential, as individual examples were not representative of overall model behavior.

## References

[1] Apostolopoulos, J. G., Chou, P. A., Culbertson, B., Kalker, T., Trott, M. D., and Wee, S., "The road to immersive communication," *Proceedings of the IEEE*, 100(4), pp. 974–990, 2012.

[2] Geronazzo, M., Tissieres, J. Y., and Serafin, S., "A minimal personalization of dynamic binaural synthesis with mixed structural modeling and scattering delay networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 411–415, IEEE, 2020.

[3] Potter, T., Cvetković, Z., and De Sena, E., "On the relative importance of visual and spatial audio rendering on VR immersion," *Frontiers in Signal Processing*, 2, p. 904866, 2022.

[4] Brinkmann, F., Aspöck, L., Ackermann, D., Lepa, S., Vorländer, M., and Weinzierl, S., "A round robin on room acoustical simulation and auralization," *The Journal of the Acoustical Society of America*, 145(4), pp. 2746–2760, 2019.

[5] Aspöck, L. and Vorländer, M., "Differences between measured and simulated room impulse responses," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 265, pp. 3209–3217, Institute of Noise Control Engineering, 2023.

[6] Pezzoli, M., Borra, F., Antonacci, F., Tubaro, S., and Sarti, A., "A parametric approach to virtual miking for sources of arbitrary directivity," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, pp. 2333–2348, 2020.

[7] Thiergart, O., Del Galdo, G., Taseska, M., and Habets, E. A., "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE transactions on audio, speech, and language processing*, 21(12), pp. 2583–2594, 2013.

[8] Pätynen, J. and Lokki, T., "Evaluation of concert hall auralization with virtual symphony orchestra," *Building Acoustics*, 18(3-4), pp. 349–366, 2011.

[9] McCarthy, R. A., Zhang, Y., Verburg, S. A., Jenkins, W. F., and Gerstoft, P., "Machine learning in acoustics: A review and open-source repository," *npj Acoustics*, 1(1), p. 18, 2025.

[10] van Waterschoot, T., "Deep, data-driven modeling of room acoustics: Literature review and research perspectives," *arXiv preprint arXiv:2504.16289*, 2025.

[11] Nichol, A. and Dhariwal, P., "Improved Denoising Diffusion Probabilistic Models," ????, doi: 10.48550/arXiv.2102.09672.

[12] Karakonstantis, X., Caviedes-Nozal, D., Richard, A., and Fernandez-Grande, E., "Room impulse response reconstruction with physics-informed deep learning," *The Journal of the Acoustical Society of America*, 155(2), pp. 1048–1059, 2024.

[13] Ratnarajah, A., Zhang, S.-X., Yu, M., Tang, Z., Manocha, D., and Yu, D., "FAST-RIR: Fast neural diffuse room impulse response generator," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 571–575, IEEE, 2022.

[14] Ratnarajah, A., Tang, Z., Aralikatti, R., and Manocha, D., "Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 924–933, 2022.

[15] Martin, I., Pastor, F., Fuentes-Hurtado, F., Belloch, J. A., Azpicueta-Ruiz, L., Naranjo, V., and Piñero, G., "Predicting room impulse responses through encoder-decoder convolutional neural networks," in *2023 IEEE 33rd international workshop on machine learning for signal processing (MLSP)*, pp. 1–6, IEEE, 2023.

[16] Song, J., Meng, C., and Ermon, S., "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[17] Jot, J.-M., "An analysis/synthesis approach to real-time artificial reverberation," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. 221–224, IEEE, 1992.

[18] Mezza, A. I., Giampiccolo, R., Bernardini, A., et al., "Differentiable Scattering Delay Networks for Artificial Reverberation," in *Proceedings of the International Conference on Digital Audio Effects (DAFx25)*, pp. 202–207, 2025.

[19] Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B., "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[20] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., et al., "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 12(1), 2016.

[21] Torre, S., Pezzoli, M., Antonacci, F., and Gannot, S., "Diffusion-RIR: Room impulse response interpolation using diffusion models," *arXiv preprint arXiv:2504.20625*, 2025.

[22] Xu, Z., Brendel, A., Prinn, A. G., and Habets, E. A., "Room Impulse Response Completion Using Signal-Prediction Diffusion Models Conditioned on Simulated Early Reflections," *arXiv preprint arXiv:2603.12442*, 2026.

[23] Vosoughi, A., Zang, Y., Yang, Q., Paek, N., Leistikow, R., and Xu, C., "PromptReverb: Multimodal Room Impulse Response Generation Through Latent Rectified Flow Matching," *arXiv preprint arXiv:2510.22439*, 2025.

[24] Siddique, N., Paheding, S., Elkin, C. P., and Devabhaktuni, V., "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE access*, 9, pp. 82031–82057, 2021.

[25] Choi, H.-S., Kim, J.-H., Huh, J., Kim, A., Ha, J.-W., and Lee, K., "PHASE-AWARE SPEECH ENHANCEMENT WITH DEEP COMPLEX U-NET," ????

[26] Natsiou, A. and O'Leary, S., "Audio representations for deep learning in sound synthesis: A review," in *2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–8, IEEE, 2021.

[27] Kong, J., Kim, J., and Bae, J., "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, 33, pp. 17022–17033, 2020.

[28] Salimans, T. and Ho, J., "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[29] Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., and Guo, B., "Efficient diffusion training via min-snr weighting strategy," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7441–7451, 2023.

[30] International Organization for Standardization, "ISO 3382-1:2009 – Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces," 2009.

[31] Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C., "Diffusion posterior sampling for general noisy inverse problems," *arXiv preprint arXiv:2209.14687*, 2022.

[32] Moliner, E., Lehtinen, J., and Välimäki, V., "Solving audio inverse problems with a diffusion model," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.