MULTI-SOURCE EXTENSION AND HYPERPARAMETER OPTIMIZATION OF THE DIFFRIR FRAMEWORK FOR ROOM IMPULSE RESPONSE SYNTHESIS

Luka Fehrmann

Dept. of Media & Digital Technology
University of Applied Sciences
St. Pölten, Austria
mp241505@ustp-students.at

Mason L. Wang

Computer Science & Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, USA
ycda@mit.edu

Martin Rumori

Dept. of Media & Digital Technology
University of Applied Sciences
St. Pölten, Austria
lbrumori@ustp.at

Peter Plessas

Dept. of Composition Studies & Music Production
University of Music and Performing Arts
Vienna, Austria
plessas@mdw.ac.at

ABSTRACT

Efficient prediction of Room Impulse Responses (RIRs) is a cornerstone for immersive virtual acoustics and scalable room acoustic modeling. This study extends the *DiffRIR* framework – proposed by Wang et al. in *Hearing Anything Anywhere* – by introducing a multi-source training logic and systematically optimizing its convergence behavior to overcome the inherent limitations of the original framework.

Our results reveal that multi-source training acts as implicit data augmentation, where the resulting increase in spatial entropy enhances the model's spectral accuracy. Furthermore, we demonstrate that the model exhibits remarkable robustness against geometric inaccuracies, maintaining numerical stability even with source positional offsets of up to 4 m in single-source baseline evaluations. By identifying a learning rate of 3×10^{-2}, we were able to reduce the training duration to 23% of the original baseline without compromising prediction accuracy.

While the increased complexity of multi-source fields necessitates a trade-off in temporal precision – quantified via our newly integrated Energy Decay Convergence (EDC) metric – this research provides an efficient and resilient solution for acoustic simulations in complex environments.

1. INTRODUCTION

The ability to accurately and efficiently synthesize Room Impulse Responses (RIRs) remains a critical challenge for the development of high-fidelity virtual acoustic environments and large-scale simulation frameworks. RIRs provide essential information for the auralization of sound fields and the estimation of acoustic room properties such as reverberation time [1]. Furthermore, they are vital for room acoustics optimization and VR applications [2].

A recent contribution in this field is *DiffRIR*[1], a hybrid differentiable room impulse response rendering framework [3]. DiffRIR bridges the gap between traditional geometric simulators and black-box models by decomposing the RIR into a deterministic part, modeling early reflections via the Image-Source Method (ISM), and a stochastic part representing late-stage reverberation through a learned, spatially invariant residual tensor [3]. Following an analysis-by-synthesis paradigm, the framework enables the gradient-based optimization of interpretable physical parameters – such as source directivity and surface reflectivity – enabling the model to generalize to unseen positions with significantly less data than purely data-driven methods [4]. Other recent developments, such as the differentiable renderer *misuka*, further underscore the efficiency of such gradient-based optimization in room acoustics [5].

DiffRIR is primarily constrained to single-source scenarios, where the acoustic field is reconstructed based on a stationary source position. A central objective of this research is to overcome this limitation: The integration of multiple distributed source positions facilitates implicit data augmentation, necessitating scalable architectures that avoid a linear increase in computational overhead [6]. Furthermore, the baseline implementation relies on conservative default hyperparameters that require extensive training durations (approx. 1,000 epochs). In the context of RIR synthesis, this represents a significant bottleneck for the scalability of large-scale acoustic modeling and the deployment of time-sensitive VR applications, where rapid model iterations are essential [3].

This research is based on previous preliminary work [7], providing a condensed evaluation and extending the framework's capabilities toward multi-source scenarios and optimized convergence behavior. In this paper, we extend the *DiffRIR* framework to address the aforementioned limitations of the original implementation. Our contributions are threefold:

1. We implement a **multi-source training logic** to verify the hypothesis that spatial diversity acts as a regularizer, enabling the model to learn from distributed source positions simultaneously within a single training cycle [8].
2. We demonstrate that a targeted adjustment of the learning rate to 3×10^{-2} enables the model to reach convergence faster, achieving a **reduction in training time** to 23% of the original baseline [9].
3. We evaluate the framework's **positional robustness** and integrate **Energy Decay Convergence (EDC)** [10] as an

[1] https://github.com/maswang32/hearinganythinganywhere/

additional metric to analyze the limits of spatially invariant late-stage modeling in complex fields.

We suggest that the inclusion of multiple source positions regularizes the learning process. By exposing the model to a higher degree of spatial information, it is assumed to decouple environment-wide physical parameters from individual source-receiver geometries, thereby improving overall spectral stability [6].

The paper is organized as follows: Section 2 provides an overview of related work. Section 3 details the implemented extensions and the evaluation metrics. Section 4 describes the evaluation setup, followed by Section 5, which presents the combined methodology and results of our experiments. Finally, Section 6 reflects on model capacity and limitations, and Section 7 concludes the paper.

2. RELATED WORK

Predicting RIRs is increasingly formulated as a high-dimensional regression task within the machine learning domain [11]. Current research predominantly follows three paradigms: purely data-driven, physics-informed, and hybrid models. Additionally, the efficiency of these models is highly dependent on optimization strategies and the choice of appropriate evaluation metrics.

2.1. Data-Driven Models

Early deep learning architectures utilized Convolutional Neural Networks (CNNs) to map the non-linear relationship between room geometry and acoustic responses [11]. More recently, *Neural Acoustic Fields* (NAF) have leveraged implicit neural representations to encode spatial coordinates into acoustic features [12]. Other approaches, such as *Mesh2IR*, utilize conditional Generative Adversarial Networks (GANs) to predict RIRs directly from 3D meshes [13]. Furthermore, recent diffusion-based generative approaches have shown significant promise in high-fidelity audio synthesis, though they often require substantial computational resources compared to hybrid models [14]. Despite their predictive accuracy, these models typically necessitate dense training datasets – often exceeding 1,000 measurements – and lack physical interpretability [4].

2.2. Physics-Informed Models

Physics-Informed Neural Networks (PINNs) integrate the wave equation or the Helmholtz equation directly into the loss function [15]. However, their application remains computationally prohibitive for high-frequency signals due to the increasing complexity of the required mesh or grid resolution.

2.3. Hybrid Approaches and DiffRIR

Hybrid approaches aim to mitigate the limitations of both purely data-driven and physics-informed models by embedding physical constraints into the learning pipeline. In this context, *DiffRIR* serves as a robust solution for the data-scarce scenarios identified in Section 2.1 [3]. By leveraging its hybrid architecture, the framework offers a computationally efficient alternative to the high-frequency wave-based simulations discussed in Section 2.2 [3, 4]. Recent frameworks like *misuka* further extend these capabilities by providing differentiable path tracing for more complex geometries [5].

2.4. Technical Foundations

The convergence efficiency of neural acoustic models is highly sensitive to hyperparameter configurations, specifically the learning rate [9]. While standard practices favor conservative learning rates to ensure numerical stability, recent work in hyperparameter optimization suggests that environment-specific "sweet spots" can drastically accelerate training without sacrificing generalization [16]. As will be shown, the learning rate is identified as a critical hyperparameter to balance the deterministic optimization of discrete early reflections against the stochastic approximation of the late-stage residual r. Furthermore, the choice of evaluation metrics is critical for assessing model quality [17]. Quantitative evaluation often relies on spectral and temporal losses like magnitude (MAG) and envelope loss (ENV) [18]. However, these metrics may fail to fully capture frequency-dependent decay characteristics [19]. Differentiable feedback delay networks (FDNs) and statistical metrics based on Schroeder backward integration, such as the Energy Decay Convergence (EDC), have been proposed to specifically evaluate the similarity and physical consistency of late-stage reverberation [10].

3. THE DIFFRIR FRAMEWORK AND EXTENSIONS

This section details the technical implementation of the foundational framework and our proposed extensions for multi-source training and hyperparameter optimization. As established in Section 1, the model synthesizes RIRs by combining a deterministic early-stage component with a stochastic late-stage residual [3].

3.1. Physical Modeling and Objective Function

Early-stage specular reflections are modeled using a differentiable implementation of the ISM [3]. To capture prominent room resonances, the ISM utilizes axial boosting for parallel surfaces up to order 50 [3]. The framework accounts for frequency-dependent source directivity via learned heatmaps and surface reflectivity through sigmoid-mapped spectral coefficients [20].

The deterministic reflections are combined with the spatially invariant late-stage residual tensor r [21]. The transition between these stages is governed by a learnable temporal spline weighting γ, resulting in the hybrid RIR formulation:

$$y(S_{xyz}, R_{xyz}) = \gamma \left[S \circledast \sum_{p \in P} K(d_p, S_p, t_p) \right] + (1 - \gamma) r \quad (1)$$

where S_{xyz} and R_{xyz} represent the source and receiver positions, respectively. S denotes the learned source response, $\circledast$ indicates convolution, and K represents the contribution of an individual reflection path $p \in P$ with outgoing direction d_p, reflecting surface sequence S_p, and time-of-arrival t_p [3]. Furthermore, K accounts for frequency-dependent air absorption and utilizes a minimum-phase transform to convert magnitude responses into the time domain [22].

Parameter optimization is achieved by minimizing a multi-scale log-spectral loss L, comparing the synthesized RIR $\hat{y}$ with ground-truth measurements y [3]. This objective function, rooted in the Differentiable Digital Signal Processing (DDSP) framework, calculates $\mathcal{L}_1$ distances of linear and logarithmic magnitude spectrograms across multiple STFT window sizes [18]. By employing

a minimum window size of 256 samples, the loss provides necessary robustness against time-of-arrival inaccuracies resulting from geometric distortions [3].

For a more in-depth explanation of the underlying *DiffRIR* framework, we refer to the original publication [3] and the associated open-source repository [8].

3.2. Multi-Source Training Logic

To enhance scalability for complex environments and address the under-determined nature of single-source optimization, we extend DiffRIR to simultaneous multi-source training without modifying the renderer of Eq. 1: a single shared parameter set (surface coefficients, late residual, spline, source response, and directivity) is jointly optimized across S source positions by replacing the single training index list with per-source index sets, one for each active source [8].

Technically, this establishes a multi-task learning framework where environment-wide parameters must explain RIRs across multiple distributed geometries within a single optimization cycle. Concretely, each per-source local microphone index is mapped to a global sample index via a block offset $i = s \cdot N_s + i_\ell$, where $s \in \{0, \ldots, S-1\}$ identifies the source, N_s is the number of receiver positions per source, and i_ℓ is the local index. The modified data loader uses this mapping to resolve i to the corresponding source-specific directory of precomputed reflection paths, ensuring that each RIR measurement is associated with its source position and reflection paths, while gradients from all sources accumulate into the shared parameter set.

Our approach resolves the physical confounding between source response and surface absorption that typically occurs under single-source supervision, specifically resolving the ambiguity between source directivity and surface reflectivity. By integrating distributed source positions, the model is exposed to a wider range of acoustic paths, which we hypothesize facilitates implicit data augmentation. This spatial diversity encourages the model to decouple shared physical parameters from source-specific geometric artifacts [6]. We evaluate two specific configurations:

- **Multi-Source-Identical:** All active sources utilize an identical set of local microphone indices. While optimizing measurement efficiency in physical setups, this configuration provides limited spatial entropy.
- **Multi-Source-Individual:** Each source is assigned a unique set of microphone positions, maximizing spatial diversity and forcing the model to learn more generalized, environment-wide acoustic features.

3.3. Evaluation Metrics

To evaluate the synthesis quality, we utilize three complementary metrics:

- **Envelope Loss (ENV):** Captures global temporal energy decay by measuring the $\mathcal{L}_1$ distance between the Hilbert-transformed envelopes of the synthesized and ground-truth RIRs [19].
- **Magnitude Loss (MAG):** Measures spectral similarity via the multiscale log-spectral $\mathcal{L}_1$ distance, capturing both transient details and steady-state spectral structures [18].
- **Energy Decay Convergence (EDC):** Frequency-selective analysis of late-stage decay using Schroeder backward integration [23]. The energy decay curve ε for a given frequency band f_c is defined as:

$$\varepsilon(t; f_c) = \sum_{\tau=t}^{L} h_{f_c}^2(\tau) \tag{2}$$

  where $h_{f_c}(\tau)$ is the RIR filtered in the respective third-octave band [10]. By evaluating the normalized squared deviation across 29 bands (20 Hz to 12.5 kHz), the EDC validates the physical consistency and statistical accuracy of the late-stage residual component r [10].

4. EVALUATION SETUP

To rigorously validate the proposed modifications, we conducted a series of experiments within a controlled acoustic environment and utilized standardized training protocols.

4.1. Database Selection and Preprocessing

The original *DiffRIR* study utilized datasets limited to single-source configurations per room, rendering them unsuitable for validating simultaneous multi-source training logic [24]. Consequently, we employ the comprehensive RIR database by Zhao et al., which provides measurements across seven diverse acoustic environments [25]. For this study, we selected the *shoe-box room* configuration ($6 \times 11.75 \times 2.8$ m) with a measured reverberation time (T_{60}) of 0.667 s [26].

This setup presents two methodological challenges: First, source and microphone positions are given as relative offsets rather than absolute coordinates [25]. This necessitates robustness evaluations regarding translational geometric shifts [4]. Second, the circular array's orientation toward the microphone zones results in a high direct-to-reverberant ratio. While beneficial for spectral accuracy (MAG), this dominance of the direct path complicates the modeling of late-stage decay (EDC) [10].

The experimental setup features a circular array of 60 loudspeakers (radius 1.5 m) and multiple planar arrays, each comprising 64 microphones [26]. We selected 12 source positions (every fifth loudspeaker, starting at 0° with 30° increments). All RIRs were standardized to a sampling rate of 48 kHz and truncated to 2.0 s (96,000 samples) to ensure compatibility with the *DiffRIR* processing pipeline and sufficient temporal depth for late-stage modeling [3]. Figure 1 illustrates the spatial arrangement and the distribution of training and validation indices.

For the six designated source positions (L01, L11, L21, L31, L41 and L51), the database provides a combined pool of 768 RIRs [26]. To evaluate the framework's performance under sparse data conditions, we conducted training using a total amount of only 24 RIRs, distributed across the active source positions to maintain high computational efficiency. This choice follows the sparse-measurement paradigm of the original framework, which demonstrated robust field recovery from as few as 12 to 24 points [3]. These training points were sampled from the pool of data points, while all remaining, non-overlapping indices were reserved for validation to ensure unbiased assessment of generalization capability.

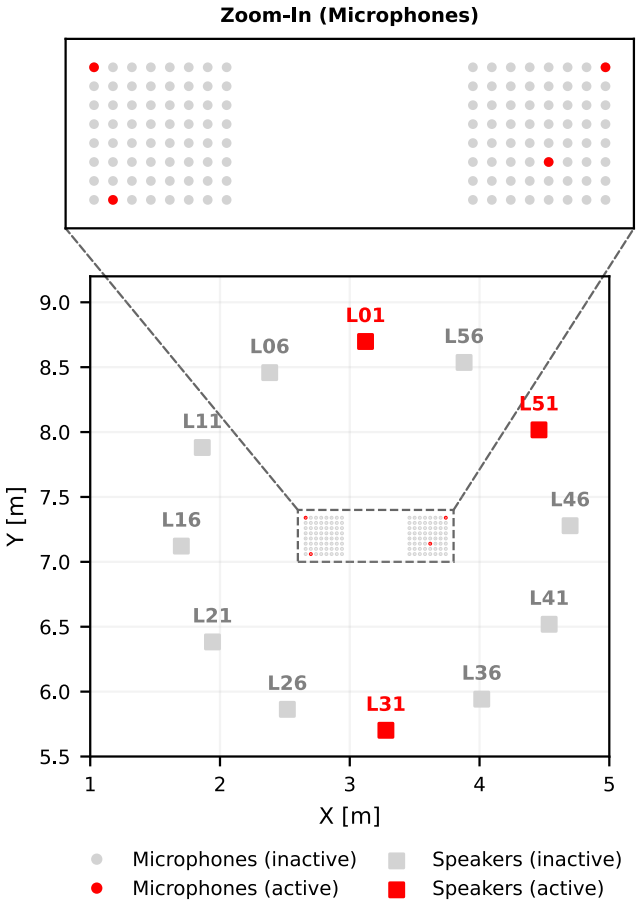

Figure 1: Spatial distribution of active training (red) and validation indices (gray) for an exemplary multi-source configuration [26]. For each of the 12 source positions, 128 RIRs from the planar microphone array were available.

4.2. Training Procedure

Training was conducted on a single NVIDIA RTX 4090 GPU. Following the established methodology of the original *DiffRIR* framework [3], results are reported from individual representative training runs. The framework's reliance on a physically-constrained, differentiable processing chain – specifically the ISM and interpretable physical parameters – inherently provides high numerical stability during optimization, which minimizes the impact of stochastic variance. For the experiments in Sections 5.1 and 5.3, the model was trained for 200 epochs using the optimized learning rate of 3×10^{-2} identified in Section 5.2. Consistent with the original framework, pink noise supervision was introduced at the halfway point of training (epoch 100) to enhance robustness against measurement noise and prevent overfitting to transient artifacts [27].

As the training duration scales with the number of RIRs in the dataset, the increased computational overhead of multi-source scenarios underscores the necessity for efficient hyperparameter selection and optimized convergence behavior [9].

5. EXPERIMENTS AND RESULTS

In this section, we evaluate the robustness, efficiency, and generalization performance of the extended *DiffRIR* framework through three sequential experiments.

5.1. Positional Robustness (Offset Study)

To assess the framework's sensitivity to geometric inaccuracies – a prevalent issue in practical scenarios where absolute coordinates are unavailable or datasets require correction [26] – we analyzed two distinct displacement scenarios:

1. **Global Translation:** Source and microphone positions were shifted by an identical vector, preserving their relative distance.
2. **Relative Displacement:** Microphone positions were shifted while the source remained stationary, thereby altering the underlying propagation paths.

The Reference configuration (0.0 m shift) corresponds to the estimated global coordinates derived from the relative positions provided in the database [26]. Given that absolute world coordinates were not explicitly provided, marginal metric fluctuations across the offset configurations are expected and demonstrate the framework's ability to maintain high synthesis quality despite such initial alignment uncertainties.

While previous experiments demonstrated robustness to mesh-based distortions (vertex shifts) of up to 1.5 m [3], we extend this analysis to translational offsets of the source and receiver positions. The metric variations reported in the original study for distortions up to 1.2,m align with our results, including similar instances where small spatial inaccuracies yielded marginally better performance than the reference configuration. This evaluation utilized a single source and 12 microphone locations, spatially distributed to maximize spatial entropy and improve acoustic field estimation.

The results compiled in Table 1 demonstrate that *DiffRIR* remains remarkable resilience to geometric shifts. Global and relative offsets of up to 4.0 m (Offsets 10 and 12) do not lead to a substantial degradation in accuracy compared to the initial configuration.

Table 1: Robustness results for various positional offsets (Metrics scaled by 10). Offsets 10 and 12 represent extreme 4.0 m shifts.

Configuration	Shift [m]	Scenario	Δ ENV	Δ MAG
Offset 6	0.2	Global	-0.03	-0.02
Reference	0.0	-	0.0	0.0
Offset 10	4.0	Global	-0.01	0.0
Offset 11	0.2	Relative	0.0	0.01
Offset 12	4.0	Relative	-0.01	-0.02

We anticipate that this stability is primarily attributed to the temporal buffering effect of the multi-scale log-spectral loss, where the 256-sample windowing allows for a sufficient region of error regarding signal onset, as well as the spatially invariant properties of the late-stage residual r, as detailed in Section 3.1. While EDC was utilized in the following experiments, it was omitted in our robustness study as the uniform nature of r ensures inherent stability of energy decay characteristics against translational shifts. Additionally, since phase information is deterministically derived via a minimum-phase transform, MAG and ENV provide a sufficient

representation of the framework's positional robustness. These characteristics provide inherent immunity to translational shifts in environments where the diffuse tail dominates the later portion of the RIR [21].

5.2. Training Efficiency and Learning Rate Optimization

Using the same single-source configuration as in the previous experiment, we evaluated the framework's sensitivity to the learning rate across two different environments. In addition to the *shoe-box room* from the database by Zhao et al. [26] we used the *classroom* configuration from the original *DiffRIR* study [24].

The baseline implementation introduced in Section 1 utilized a conservative learning rate of 1×10^{-4}, requiring approximately 1,000 epochs to reach a stable loss minimum [3]. On our utilized hardware, this corresponds to a total training duration of 10 h 20 min, including initialization and the calculation of evaluation metrics. Our empirical evaluation identified 3×10^{-2} as the optimal learning rate for both configurations, accelerating the optimization process as illustrated in Fig. 2.

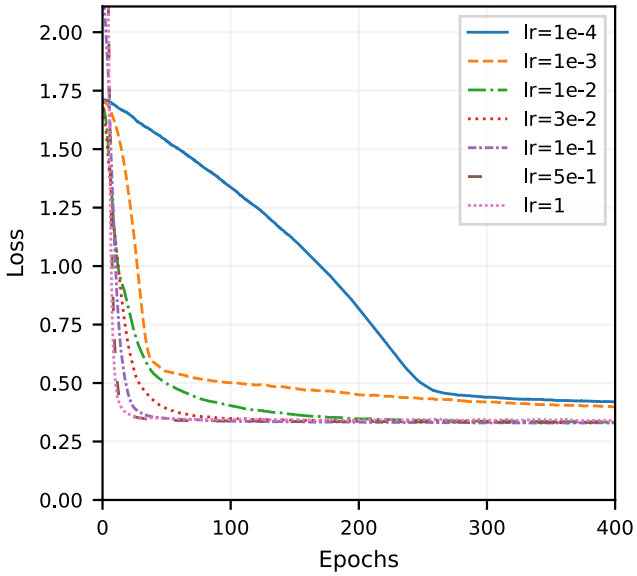

Figure 2: Training loss convergence for various learning rates (10^{-4} to 1) utilizing the *shoe-box* environment from the database by Zhao et al. [26]. The optimized rate of 3×10^{-2} achieves a stable minimum within 80 to 100 epochs, noticeably accelerating the optimization process compared to the single-source baseline.

As summarized in Table 2, this hyperparameter optimization enables a reduction of training epochs from 1,000 to 200, effectively reducing the total training duration to 2 h 20 min – approximately 23% of the original baseline. While the *classroom* configuration exhibits minor fluctuations across metrics – including a marginal increase in spectral error (MAG +0.08) alongside slight improvements in temporal envelope (ENV -0.04) and decay consistency (EDC -0.03) – the *shoe-box* environment demonstrates a substantial performance gain across all evaluated metrics. Specifically, MAG in the *shoe-box* room improved from 2.86 to 2.21, while the EDC error was more than halved.

Despite these improvements, a comparison reveals that the EDC metric remains significantly better for the *classroom* dataset

Table 2: Efficiency Comparison: Baseline vs. Optimized Learning Rate (LR) across Datasets (Metrics scaled by 10).

Dataset	Configuration (LR)	Epochs	ENV	MAG	EDC
Classroom	Baseline (1×10^{-4})	1,000	1.00	5.22	0.13
	Optimized (3×10^{-2})	**200**	**0.96**	**5.30**	**0.10**
Shoe-box	Baseline (1×10^{-4})	1,000	0.44	2.86	6.15
	Optimized (3×10^{-2})	**200**	**0.31**	**2.21**	**3.06**

(0.10 vs. 3.06). This disparity is likely attributable to the high direct-to-reverberant ratio in the *shoe-box room*, which possibly complicates the modeling of late-stage decay characteristics compared to the more diffuse *classroom* environment [26].

Analysis suggests that the late-stage residual r specifically benefits from higher learning rates, evidenced by the sustained reduction in EDC error (see Fig. 3). To account for the slower convergence observed at smaller learning rates, as shown in Fig. 2, metrics were evaluated after 500 epochs, ensuring that all configurations have reached a stable minimum where no further significant loss reduction is expected. While EDC continues to improve at even higher rates, the simultaneous degradation of MAG and ENV identifies 3×10^{-2} as the optimal balance between speed and generalization.

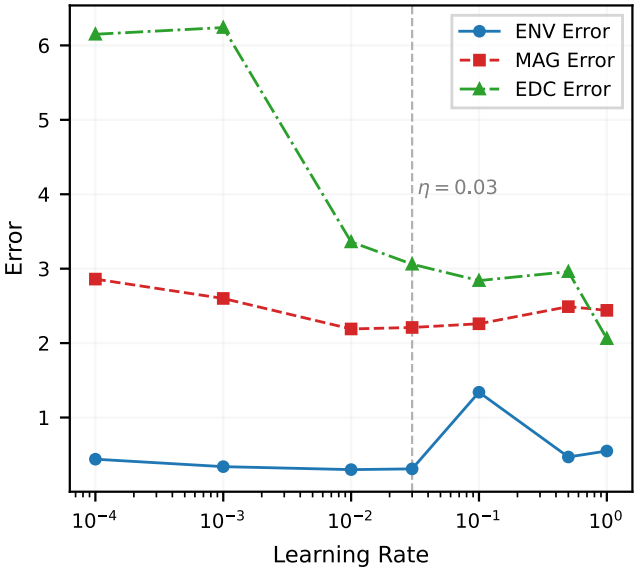

Figure 3: Validation metrics (ENV, MAG, EDC) as a function of the learning rate, measured after 500 training epochs for the *shoe box* configuration. The x-axis is displayed on a logarithmic scale, and a learning rate of 3×10^{-2} represents the optimal balance. The y-axis denotes the individual error scores (scaled by 10) for each metric.

5.3. Multi-Source Performance and Regularization

A central objective of this study was to evaluate the framework's capability to learn from multiple source positions simultaneously. We investigated the two data selection strategies, Multi-Source-Identical and Multi-Source-Individual, maintaining a constant total

training volume $N \leq 24$ to ensure a rigorous comparison with single-source baselines.

As shown in Table 3, multi-source training yields a consistent improvement in spectral accuracy across all tested configurations. Notably, every multi-source setup outperformed the corresponding single-source baseline in the MAG metric across all evaluated training set sizes N, demonstrating a robust and consistent performance trend. For $N = 24$, the Multi-Source-Individual approach achieved the lowest MAG of 2.10, compared to 2.31 in the single-source baseline. The magnitude of this improvement (0.21) is comparable to the performance impact of central model components reported in the original framework's ablation studies, thereby contextualizing its relevance within the DiffRIR architecture [3]. Our results demonstrate a substantial compensatory effect: a higher number of distributed source positions can effectively compensate for a limited number of microphones per source, provided the total data volume is sufficient. These results empirically validate the hypothesis from Section 3.2, confirming that increased spatial variability acts as implicit data augmentation, providing a regularization effect that forces the model to decouple shared physical parameters, such as surface materials, from source-specific geometric artifacts [6].

Table 3: Comparison of Single-Source and Multi-Source (MS) training across different training set sizes N (Metrics scaled by 10). Multi-source configurations consistently yield superior MAG values.

	Single-Source			MS-Identical			MS-Individual		
N	MAG	ENV	EDC	MAG	ENV	EDC	MAG	ENV	EDC
6	2.36	0.33	4.30	2.25	0.30	4.40	2.24	0.31	4.44
12	2.30	0.33	3.90	2.16	0.30	4.01	2.16	0.31	3.95
24	2.31	**0.25**	**3.68**	2.11	0.31	3.80	**2.10**	0.30	3.78

However, the frequency-selective EDC analysis reveals a fundamental performance trade-off. While spectral accuracy (MAG) and temporal envelopes (ENV) benefit from the increased spatial entropy, multi-source configurations consistently exhibit slightly higher EDC errors than single-source baselines. Specifically, the absolute EDC error for multi-source setups decreases from 4.44 at $N = 6$ to 3.78 at $N = 24$, yet it remains systematically higher than the single-source baseline of 3.68 at an equivalent volume. The systematic nature of this gap across all training volumes indicates a structural trade-off, consistent with the heightened sensitivity of EDC-based metrics to physical reverberation features as analyzed by Dal Santo et al. [10]. This relative performance gap provides empirical support for the capacity limit hypothesis discussed in Section 6.1.

6. DISCUSSION

The experimental results, particularly the observed trade-off between spectral gains (MAG) and reverberation decay precision (EDC) in multi-source scenarios, necessitate a critical reflection on the underlying physical assumptions and model constraints.

6.1. Isotropy and Model Capacity

While *DiffRIR* relies on the idealization of isotropic late-stage reverberation, real-world environments like the *shoe-box room* often deviate from this uniform decay due to non-uniform absorption patterns. As Nolan et al. demonstrate, such patterns favor the separation of fast-decaying non-grazing waves from slow-decaying grazing waves traveling parallel to reflective surfaces, which prevents the establishment of an ideally diffuse sound field [21]. In the utilized database provided by Zhao et al., this lack of perfect isotropy – further amplified by direct-sound dominance resulting from the circular speaker array's orientation toward the microphone zones – introduces a crucial strain on model capacity when approximating multiple late-fields simultaneously [26].

While a single spatially invariant tensor r is sufficient to approximate these characteristics for a single source, simultaneous multi-source training places a substantial demand on model capacity. Each additional source requires the optimization of individual parameters for directivity and source response, increasing the complexity of the global optimization task. This transition effectively addresses the structurally non-identifiable nature of the single-source inverse problem by imposing a physical consistency constraint across multiple geometries.

The hypothesis of a capacity limit is further supported by the persistent relative performance gap observed at identical training volumes. Although absolute EDC errors decrease as more sources are added – e.g., from 4.44 at $N = 6$ to 3.78 at $N = 24$ – the multi-source setups consistently fail to match the temporal precision of the single-source baseline (3.68 at $N = 24$). This relative disparity suggests that the spatially invariant residual r reaches a structural bottleneck when forced to represent multiple, source-dependent late-fields that do not conform to the ideal isotropy assumption [21].

This imbalance in internal resource allocation explains why the model prioritizes spectral accuracy through implicit regularization while incurring minor losses in the temporal precision of the energy decay. Furthermore, the inclusion of overlapping microphone positions across different source configurations likely provides additional structural learning cues which, following the acoustic machine learning paradigm by Bianco et al., enables the extraction of latent representations from redundant spatial information [11]. By increasing spatial entropy through distributed source positions, the model identifies structural acoustic features that are invariant to specific source positions, resolving the confounding of environment-wide features and source-specific geometric biases and leading to the observed superior spectral accuracy (MAG) compared to single-source setups. This strategy effectively acts as geometric data augmentation, as observed by Taylor and Nitschke, forcing the network to learn generalized environment features through multi-task regularization [6].

6.2. Positional Robustness and Geometric Invariances

Despite internal capacity constraints observed in multi-source scenarios, the framework's remarkable robustness to positional offsets of up to 4.0 m demonstrates that the late-stage residual effectively captures the global statistical properties of the acoustic space. This stability is primarily attributed to the multi-scale log-spectral loss, where the 256-sample STFT windowing acts as a temporal buffer. Consistent with the original study, this minimum window size compensates for time-of-arrival inaccuracies by ensuring that the spectral energy of a reflection remains captured within the same analysis window, even if geometric discrepancies cause small misalignments [3].

Furthermore, the spatially invariant late-stage residual r provides inherent immunity to translational shifts by approximating the late-field as a statistically uniform and isotropic sound field.

Following Nolan et al.'s definition of such fields as being homogeneous, the global statistical properties of the modeled late-field remain invariant to the observer's position, providing stability as long as the relative configuration between source and receiver is maintained [21].

Our findings further suggest that the framework is more resilient to such translational offsets than to geometric distortions: In simple geometries, shifting positions primarily affects path lengths without altering the topology of early reflection paths. In contrast, distorting room geometry increases acoustic complexity, making the deterministic determination of paths via the ISM more challenging.

7. CONCLUSION

This study presented a systematic evaluation and extension of the *DiffRIR* framework, identifying key factors for its efficient and robust application in complex environments. Based on previous preliminary work [7], we demonstrated that an optimized learning rate of 3×10^{-2} accelerates convergence to 23% of the original baseline without compromising accuracy. Furthermore, the framework exhibits remarkable resilience to geometric inaccuracies, maintaining stable performance even with positional offsets of up to 4.0 m [3]. These findings are critical for large-scale acoustic simulations where computational efficiency and robustness are primary constraints.

Our results indicate that multi-source training acts as a form of implicit data augmentation, offering a promising approach to enhance spectral accuracy through increased spatial entropy [6]. By increasing spatial entropy through distributed source positions, the model identifies structural acoustic features that are invariant to specific source positions, leading to superior spectral accuracy (MAG) compared to single-source setups. For practical deployments, we recommend maximizing spatial variability through individual microphone positions to ensure robust physical parameter estimation.

Frequency-selective EDC analysis, however, revealed a fundamental trade-off, suggesting that the spatially invariant late-stage residual r might reach its capacity limit when representing multiple, complex reverberation fields simultaneously [21]. Future research should address these limits by investigating source-dependent or hierarchical residual structures [12]. Along these lines, future investigations into error decomposition between the residual and deterministic components – complemented by reflection path visualizations – could further clarify the accuracy of the optimized physical parameters. Additionally, while this study serves as a proof-of-concept for the multi-source logic, extending evaluations to a broader range of acoustic environments and conducting additional ablation studies is planned to further strengthen the conclusions regarding model generalization. Furthermore, comparing the framework's efficiency against emerging diffusion-based generative models is an important direction for future evaluation. Finally, subjective listening tests and the application of perceptual metrics for coloration are required to validate the perceptual relevance of the observed spectral gains in immersive audio environments.

8. ACKNOWLEDGMENTS

The authors would like to thank Sipei Zhao from the University of Technology Sydney for providing the RIR database used in this study. Furthermore, we would like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve this manuscript.

9. DECLARATION OF AI USE

The authors used Google's Gemini 3 Flash for linguistic polishing, including improvements to grammar, clarity, and structure. Following these suggestions, the authors manually reviewed and refined the manuscript and take full responsibility for the final content.

10. REFERENCES

[1] A. Kujawski, A. J. Pelling, and E. Sarradj, "Miracle—a microphone array impulse response dataset for acoustic learning," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 32, 2024.

[2] J.-W. Lo, Y.-S. Chen, and M. R. Bai, "Physics-informed machine learning for sound field estimation: fundamentals, state of the art, and challenges," *npj Acoustics*, 2026, downloaded from academic.oup.com, March 2026.

[3] M. L. Wang, R. Sawata, S. Clarke, R. Gao, S. Wu, and J. Wu, "Hearing Anything Anywhere," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 790–11 799.

[4] M. Pezzoli, D. Perini, A. Bernardini, F. Borra, F. Antonacci, and A. Sarti, "Deep prior approach for room impulse response reconstruction," *Sensors*, vol. 22, no. 7, p. 2710, 2022.

[5] T. Jüterbock, U. Finnendahl, M. Worchel, D. Wujecki, M. Alexa, and S. Weinzierl, "misuka: An open-source differentiable room acoustic renderer," in *Proceedings of Meetings on Acoustics*, vol. 58, no. 022004, 2025.

[6] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *2018 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2018, pp. 1542–1547.

[7] L. Fehrmann, "Evaluierung und Erweiterung eines differenzierbaren maschinellen Lernmodells zur Synthese von Raumimpulsantworten," Masterarbeit, Fachhochschule St. Pölten, St. Pölten, Austria, September 2025.

[8] M. L. Wang, R. Sawata, S. Clarke, R. Gao, S. Wu, and J. Wu, "hearinganythinganywhere," june 2024. [Online]. Available: https://github.com/maswang32/hearinganythinganywhere/

[9] M. A. K. Raiaan, S. Sakib, N. M. Fahad, A. Al Mamun, M. A. Rahman, S. Shatabda, and M. S. H. Mukta, "A systematic review of hyperparameter optimization techniques in convolutional neural networks," *Decision Analytics Journal*, vol. 11, p. 100470, 2024.

[10] G. Dal Santo, K. Prawda, S. J. Schlecht, and V. Välimäki, "Similarity metrics for late reverberation," in *2024 58th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2024, pp. 1409–1413.

[11] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," *The Journal of the Acoustical Society of America*, vol. 146, no. 5, pp. 3590–3628, 2019.

[12] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3165–3177, 2022.

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

[14] M. Wang and C.-Z. A. Huang, "Latent fourier transform," in *The Fourteenth International Conference on Learning Representations*, 2026.

[15] X. Karakonstantis, D. Caviedes-Nozal, A. Richard, and E. Fernandez-Grande, "Room impulse response reconstruction with physics-informed deep learning," *The Journal of the Acoustical Society of America*, vol. 155, no. 2, pp. 1048–1059, 2024.

[16] T. Yu and H. Zhu, "Hyper-parameter optimization: A review of algorithms and applications," 2020. [Online]. Available: https://arxiv.org/abs/2003.05689

[17] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, p. 6086, 2024.

[18] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," *arXiv preprint arXiv:2001.04643*, 2020.

[19] S. De Cesaris, D. D'Orazio, F. Morandi, and M. Garai, "Extraction of the envelope from impulse responses using preprocessed energy detection for early decay estimation," *The Journal of the Acoustical Society of America*, vol. 138, no. 4, pp. 2513–2523, 2015.

[20] K. Zhang, F. Luan, Q. Wang, K. Bala, and N. Snavely, "Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5453–5462.

[21] M. Nolan, M. Berzborn, and E. Fernandez-Grande, "Isotropy in decaying reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 148, no. 2, pp. 1077–1088, 2020.

[22] J. G. McDaniel and C. L. Clarke, "Interpretation and identification of minimum phase reflection coefficients," *The Journal of the Acoustical Society of America*, vol. 110, no. 6, pp. 3003–3010, 2001.

[23] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[24] M. L. Wang, R. Sawata, S. Clarke, R. Gao, S. Wu, and J. Wu, "Hearing Anything Anywhere - the DIFFRIR dataset," May 2024, accessed: 2025-03-15. [Online]. Available: https://doi.org/10.5281/zenodo.11195833

[25] S. Zhao, Q. Zhu, E. Cheng, and I. S. Burnett, "A room impulse response database for multizone sound field reproduction," 2022, accessed: 2025-05-02. [Online]. Available: https://doi.org/10.26195/0wx8-v473

[26] S. Zhao, Q. Zhu, E. Cheng, and I. S. Burnett, "A room impulse response database for multizone sound field reproduction (L)," *The Journal of the Acoustical Society of America*, vol. 152, no. 4, pp. 2505–2512, 2022.

[27] A. Richard, P. Dodds, and V. K. Ithapu, "Deep impulse responses: Estimating and parameterizing filters with deep networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3209–3213.