MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Echoes after Edits: Room Impulse Response Estimation for Geometry Update

Bhosale, Swapnil; Masuyama, Yoshiki; Chatterjee, Moitreya; Boeddeker, Christoph; Richter, Julius; Wichern, Gordon; Le Roux, Jonathan

TR2026-140 September 29, 2026

Abstract

Room impulse responses (RIRs) characterize sound propagation in a scene, determined by its geometry and materials. Traditional RIR estimation relies on acoustic simulation, while recent deep learning focuses on spatial interpolation. However, acoustic simulation requires material annotation of each object, and interpolation requires multiple RIRs whenever the scene is edited slightly (e.g., rearranging furniture). Both approaches are inefficient and impractical in everyday environments. In this paper, we introduce edit-conditioned RIR estimation, which updates a measured RIR to reflect geometry edits. To relate the geometry edits to acoustic variation without material annotations, we simulate proxy RIRs from the room meshes before and after the edit using predefined materials. We propose PG-RIR, which leverages these proxies to estimate the variation in the real RIR. Experiments show that PG-RIR accurately predicts RIRs under furniture rearrangement and removal.

Interspeech 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Echoes after Edits: Room Impulse Response Estimation for Geometry Update

Swapnil Bhosale[1,2], *Yoshiki Masuyama*[1], *Moitreya Chatterjee*[1], *Christoph Boeddeker*[1], *Julius Richter*[1], *Gordon Wichern*[1], *Jonathan Le Roux*[1]

[1] Mitsubishi Electric Research Laboratories (MERL), Cambridge, USA
[2] University of Surrey, UK

s.bhosale@surrey.ac.uk, {masuyama,chatterjee,boeddeker,richter,wichern,leroux}@merl.com

Abstract

Room impulse responses (RIRs) characterize sound propagation in a scene, determined by its geometry and materials. Traditional RIR estimation relies on acoustic simulation, while recent deep learning focuses on spatial interpolation. However, acoustic simulation requires material annotation of each object, and interpolation requires multiple RIRs whenever the scene is edited slightly (e.g., rearranging furniture). Both approaches are inefficient and impractical in everyday environments. In this paper, we introduce edit-conditioned RIR estimation, which updates a measured RIR to reflect geometry edits. To relate the geometry edits to acoustic variation without material annotations, we simulate proxy RIRs from the room meshes before and after the edit using predefined materials. We propose `PG-RIR`, which leverages these proxies to estimate the variation in the real RIR. Experiments show that `PG-RIR` accurately predicts RIRs under furniture rearrangement and removal.

Index Terms: room impulse response, neural acoustic modeling, geometry edits.

1. Introduction

Room impulse responses (RIR) characterize sound propagation in an indoor environment and determine distinctive auditory perception in each environment [1, 2]. RIRs capture not only the direct sound but also early reflections and late reverberation, acting as a compact signature of how a scene's geometry and surface materials shape sound propagation.

Prior work in room acoustics has mainly focused on two tasks: RIR simulation [3–6] and sound field interpolation [7–9]. Acoustic simulation can produce physically-precise RIRs given fine-grained 3D meshes and surface properties (e.g., absorption coefficients). These conditions are, however, not easy to acquire at scale in real-world environments. Sound field interpolation has recently advanced with neural fields (NFs) [10–15], which attempt to learn a continuous map from source-receiver positions to RIRs. NFs are typically trained for each scene separately, demanding dense RIR measurements. To circumvent this expensive measurement process, cross-scene RIR modeling has emerged [16–19]; this approach predicts RIRs within a new scene from a few reference RIRs and/or visual cues.

Despite these recent advances, spatial interpolation is fundamentally restricted to a static scene and requires a small set of reference RIRs for every scene. This assumption is highly restrictive in practical situations, where the indoor environment can be edited frequently (e.g., rearrangement and removal of furniture/objects). Even small geometric edits can produce significant changes in the acoustic reflection and scattering, necessitating re-measurement of RIRs [20–22].

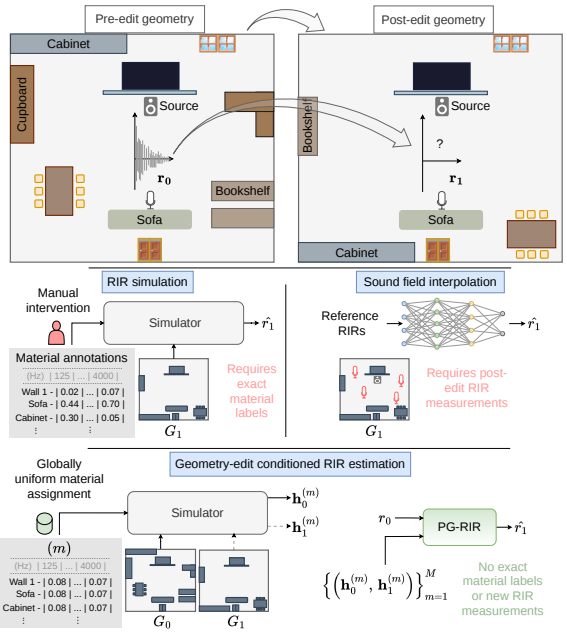

Figure 1: ***RIR estimation under scene edits.*** *Top: When a scene changes ($G_0 \rightarrow G_1$), the RIR varies from $\mathbf{r}_0$ to $\mathbf{r}_1$. Middle left: Acoustic simulation requires exhaustive human-annotated material coefficients for the edited scene. Middle right: Spatial interpolation requires capturing multiple reference RIR measurements inside G_1. Bottom: Our proposed `PG-RIR` predicts the target response $\hat{\mathbf{r}}_1$ zero-shot by using the pre-edit observation $\mathbf{r}_0$ and homogeneous proxies simulated from the geometries.*

This motivates us to explore another task, RIR estimation under scene edits as illustrated in the top panel of Fig. 1. In particular, focusing on geometric edits, we aim to estimate the RIR $\mathbf{r}_1$ measured after the edit from a prior RIR $\mathbf{r}_0$ at the same source-receiver positions, by leveraging a representation of room geometry before and after the edits G_0 and G_1. This task must be addressed without a time-consuming acoustic measurement process, such as capturing new RIRs in G_1 or obtaining fine-grained material annotation for each object. Given that acquiring updated 3D meshes (G_0, G_1) is now highly practical using ubiquitous consumer depth sensors [23, 24], a straightforward approach is to train a neural network that directly predicts $\mathbf{r}_1$ from the triplet $(\mathbf{r}_0, G_0, G_1)$. However, such an approach would require enormous amounts of training data to learn the complicated physics of sound propagation.

This work was done while S. Bhosale was an intern at MERL.

To mitigate this issue, we propose to leverage RIR simulation with predefined materials as a proxy for G_0 and G_1. Specifically, we simulate RIRs using an identical, globally uniform surface material, describing how the geometric edits affect sound propagation without accessing scene-specific fine-grained material properties. The difference between the proxy RIRs for G_0 and G_1 provides an interpretable, physics-grounded cue to estimate $\mathbf{r}_1$. We then develop PG-RIR, an RIR estimation network that predicts the difference between $\mathbf{r}_1$ and $\mathbf{r}_0$ given $\mathbf{r}_0$ and a set of proxy RIRs. In our experiments on simulated RIRs, PG-RIR achieves accurate estimation of $\mathbf{r}_1$. Furthermore, PG-RIR generalizes well to furniture-rearrangement situations while it is trained only on furniture removals.

2. Methodology

2.1. Problem Formulation

Let G_0 denote a representation of room geometry and $\mathbf{r}_0 \in \mathbb{R}^L$ denote the measured RIR for a given source and receiver pair in G_0, where L is the number of samples. We consider a scenario where the scene undergoes edits (e.g., object removals or translations), leading to geometry G_1 and a corresponding target response $\mathbf{r}_1 \in \mathbb{R}^L$ for the same source and receiver pair. Our objective is to predict $\mathbf{r}_1$ from $\mathbf{r}_0$ without requiring access to explicit surface material annotations, while only assuming access to the room geometry (3D mesh) before and after the edit. We formulate this task as learning a parametric mapping:

$$\hat{\mathbf{r}}_1 = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{r}_0, \mathcal{P}_{G_0}, \mathcal{P}_{G_1}), \tag{1}$$

where $\mathcal{P}_G$ is an RIR-related encoding of the pre-edit (G_0) or post-edit (G_1) room geometry, and $\boldsymbol{\theta}$ denotes the model parameters. We expect that the model can combine the information about geometry-induced acoustic variation obtained from $\mathcal{P}_{G_0}$ and $\mathcal{P}_{G_1}$ with fine-grained material information implicitly captured in $\mathbf{r}_0$ in order to perform this transformation.

Given a geometry G_n, we utilize an established acoustic simulation to obtain $\mathcal{P}_{G_n}$. Specifically, for each geometry, we simulate M RIRs at the same source-receiver position with M predefined materials:

$$\mathcal{P}_{G_n} = (\mathbf{h}_n^{(1)}, \ldots, \mathbf{h}_n^{(M)}), \tag{2}$$

where $n \in \{0, 1\}$ is the geometry index, and $\mathbf{h}_n^{(m)} \in \mathbb{R}^L$ denotes the simulated RIR in G_n with a predefined global material c_m. As $\mathbf{h}_0^{(m)}$ and $\mathbf{h}_1^{(m)}$ are computed with identical materials, their difference provides information about the geometry-induced acoustic variation under controlled absorption regimes.

2.2. PG-RIR

We propose PG-RIR, a proxy-guided RIR estimation network, to predict the scene-edit-induced RIR transformation from $\mathbf{r}_0$ to $\mathbf{r}_1$ in the time-frequency domain. As illustrated in Fig. 2, the proxy RIRs are converted to a log-magnitude spectrogram $\mathbf{X}_n^{(m)} \in \mathbb{R}^{F \times T}$ using the short-time Fourier transform (STFT), where F and T are the number of frequency bins and time frames, respectively. We denote the log-magnitude spectrogram of the real pre-edit and post-edit RIRs ($\mathbf{r}_0$ and $\mathbf{r}_1$) as $\mathbf{X}_0^{(0)}$ and $\mathbf{X}_1^{(0)}$, respectively. Following Liu et al. [25], each spectrogram is encoded to an acoustic feature using a shared ResNet-18 [26] encoder. Next, we introduce trainable *type embeddings* (indicating their association with the real pre-edit response, a pre-edit proxy, or a post-edit proxy) and *material embeddings* (indicating the specific proxy absorption m). The sum of the corresponding embeddings is concatenated with each acoustic feature and passed to a projection layer, resulting in fused latents $\mathbf{z}_n^{(m)} \in \mathbb{R}^C$, where C is the embedding size.

Context-Aware Queries: To summarize the geometry edit, we pool the fused latents $\mathbf{z}_n^{(m)}$ by their input type. Specifically, the context-aware query is generated by concatenating type-wise embeddings and passing them to a multi-layer perceptron ϕ:

$$\mathbf{q} = \phi\Big(\big[\mathbf{z}_0^{(0)\mathsf{T}}, \frac{1}{M}\sum_{m=1}^{M}\mathbf{z}_0^{(m)\mathsf{T}}, \frac{1}{M}\sum_{m=1}^{M}\mathbf{z}_1^{(m)\mathsf{T}}\big]^{\mathsf{T}}\Big). \tag{3}$$

We combine the reference RIRs using relevance weights,

$$\alpha_n^{(m)} = \frac{\exp\left(\langle \mathbf{q}, \mathbf{z}_n^{(m)}\rangle/\sqrt{C}\right)}{\sum_{n',m'}\exp\left(\langle \mathbf{q}, \mathbf{z}_{n'}^{(m')}\rangle/\sqrt{C}\right)}, \tag{4}$$

where $\langle\cdot,\cdot\rangle$ is the inner product.

To model non-linear reverberation decay, we utilize a query-modulated hypernetwork [27]. To emphasize important acoustic information in the lower frequencies, we introduce B frequency bands mapped to the high-resolution STFT dimension F using a base-2 octave progression $b(f)$. We compute sinusoidal positional encodings of the normalized time steps [25] and project them at each time t and for each frequency band b into $\beta(b,t) \in \mathbb{R}^C$. Again for each band b, we use the target query to generate a gating vector $g(b,\mathbf{q}) = \sigma(\mathbf{W}_g(b)\mathbf{q}) \in \mathbb{R}^C$, where $\mathbf{W}_g(b) \in \mathbb{R}^{C\times C}$ for any b, from which we compute a dynamically scaled temporal basis $\tilde{\beta}(b,t;\mathbf{q}) = \beta(b,t) \odot g(b,\mathbf{q}) \in \mathbb{R}^C$. For each reference (n,m) and frequency band b, we compute temporal modulation weights as $w_n^{(m)}(b,t) = \tilde{\beta}(b,t;\mathbf{q})^\top \mathbf{z}_n^{(m)}$. The predicted residual log-spectrogram is then formed by weighting the reference spectrograms (including $\mathbf{X}_0^{(0)}$):

$$\Delta\hat{\mathbf{X}}(f,t) = \sum_{n,m}\alpha_n^{(m)}\, w_n^{(m)}(b(f),t)\, \mathbf{X}_n^{(m)}(f,t). \tag{5}$$

We compute the final estimate as $\hat{\mathbf{X}}_1^{(0)} = \mathbf{X}_0^{(0)} + \Delta\hat{\mathbf{X}}$, which is then converted to the time domain by inverse STFT with the phase of $\mathbf{r}_0$. PG-RIR shown in Fig. 2 has 11.9 M parameters.

2.3. Learning Objectives

We train PG-RIR using a sum of two common loss functions on the log-magnitude spectra of the predicted RIR $\hat{\mathbf{X}}_1^{(0)}$ and the target RIR $\mathbf{X}_1^{(0)}$:

$$\mathcal{L} = \lambda_{\mathrm{s}}\big\|\hat{\mathbf{X}}_1^{(0)} - \mathbf{X}_1^{(0)}\big\|_1 + \lambda_{\mathrm{d}}\big\|\hat{\mathbf{E}}_1^{(0)} - \mathbf{E}_1^{(0)}\big\|_1, \tag{6}$$

$$E_n^{(0)}(f,t) = 10\log_{10}\left(\sum_{\tau=t}^{T}\exp(\mathbf{X}_n^{(0)}(f,\tau))^2\right), \tag{7}$$

where $\lambda_{\mathrm{s}} = 1$ and $\lambda_{\mathrm{d}} = 0.1$ are weights for the spectral and decay losses. The first term of (6) enforces fine-grained fidelity to $\mathbf{X}_1^{(0)}$ in the time-frequency domain. Meanwhile, the second term is based on the band-wise energy decay curves [28], which enforces the natural decay of the RIR. We approximate the decay in the STFT domain as in (7).

3. Dataset

A core challenge in learning geometry-induced acoustic variation is the lack of datasets and hence we construct a large-scale dataset of paired RIRs simulated before and after controlled geometry edits.

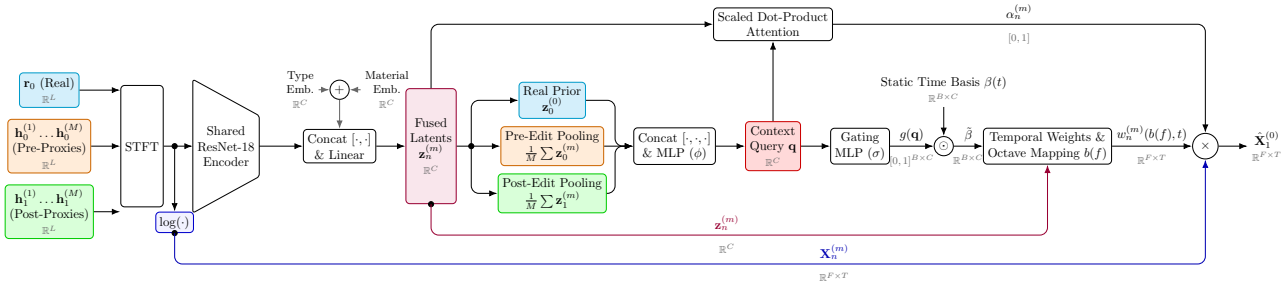

Figure 2: ***PG-RIR.*** *An encoder extracts features from the reference and proxy responses, fusing them with type and material embeddings into structured latents* $\mathbf{Z}$*. Type-specific pooling generates a context query* $\mathbf{q}$ *that guides global attention (*$\alpha_n^{(m)}$*) and modulates (*$\otimes$*) a temporal basis. The resulting weights modulate the reference log-spectrograms to synthesize the acoustic residual* $\Delta\hat{\mathbf{X}}$*.*

Scene edits: We use 14 indoor environments curated from the iGibson dataset [29], following [30], as the original scenes G_0. These scenes cover a diverse distribution of room scales and types (e.g., bedrooms, large living spaces, and kitchens). For each scene, we sample physically valid source and receiver positions by voxelizing the navigable free space. In our training and primary evaluation sets, we sample multiple objects from G_0 and remove them to obtain G_1. We also construct an additional evaluation set with *translation* edits, where objects in G_0 are rearranged, to validate the generalization capability of PG-RIR under unseen edits. We restrict edits to avoid changes to the background geometry (e.g., walls, ceilings), producing localized geometric edits.

Acoustic Simulation: For each sampled source and receiver pair in an edit tuple (G_0, G_1), we render ground-truth RIRs $(\mathbf{r}_0, \mathbf{r}_1)$ using AcoustiX simulator [30] with the realistic, object-wise material properties provided by the iGibson assets. In addition, we simulate M proxy RIRs by overriding all surfaces with M different global material parameters. AcoustiX is based on ray tracing and synthesizes RIRs by tracking energy paths and assigning stochastic phase shifts (e.g., random sign flips). This randomness prevents learning structured residuals. To mitigate this randomness, we enforce phase consistency across acoustic simulations in G_0 and G_1 by determining phase assignment based on geometric properties[1] (e.g., path length, coordinates, reflection angles).

Dataset Statistics: To force the model to learn the relationship between the geometry edit and the acoustic variation, we discard source–receiver pairs with reverberation-time changes $\Delta T_{60} > 0.6$ s or clarity changes $\Delta C_{50} > 10$ dB. The final dataset contains 143,455 valid source–receiver pairs across 14 base geometries (G_0) and 264 edited variants (G_1). For analysis across edit magnitudes, we stratify samples into four tiers by ground-truth ΔT_{60}. Typical domestic environments have base T_{60} values of 0.2 to 0.5 s. We thus define: Tier 1 (High): $\Delta T_{60} > 0.08$ s (often over 20% relative change); Tier 2 (Moderate): $0.05 < \Delta T_{60} \leq 0.08$ s; Tier 3 (Subtle): $0.03 < \Delta T_{60} \leq 0.05$ s; Tier 4 (Static): $\Delta T_{60} \leq 0.03$ s. These tiers account for 26%, 13%, 21%, and 40% of the dataset, respectively. This tiering provides an interpretable axis of geometry-induced reverberation change, enabling targeted evaluation of the model's sensitivity to both subtle and substantial edits.

4. Experiments

Experiments are conducted across 14 scenes using a leave-one-scene-out protocol, reporting average performance on the held-out runs. We evaluate the transformations by reporting standard room-acoustic metrics [12, 15, 19] computed between the predicted $\hat{\mathbf{r}}_1$ and ground-truth $\mathbf{r}_1$: reverberation time (T_{60}), early decay time (EDT), and clarity (C_{50}).

[1] https://github.com/merlresearch/geometry-edit-rir

Table 1: *Leave-one-scene-out evaluation across 14 unseen scenes. Mean ± standard deviation of absolute errors for each metric. Lower is better.*

Method	$T_{60} \downarrow$	$C_{50} \downarrow$	EDT $\downarrow$
Identity ($\mathbf{r}_0$)	0.0432 ± 0.0073	2.034 ± 0.518	0.032 ± 0.0056
xRIR (K=4)	0.0523 ± 0.0053	3.548 ± 0.414	0.045 ± 0.0079
xRIR (K=8)	0.0538 ± 0.0058	3.147 ± 0.493	0.036 ± 0.0063
PG-RIR	$\mathbf{0.0316 \pm 0.0076}$	$\mathbf{1.407 \pm 0.435}$	$\mathbf{0.022 \pm 0.0044}$

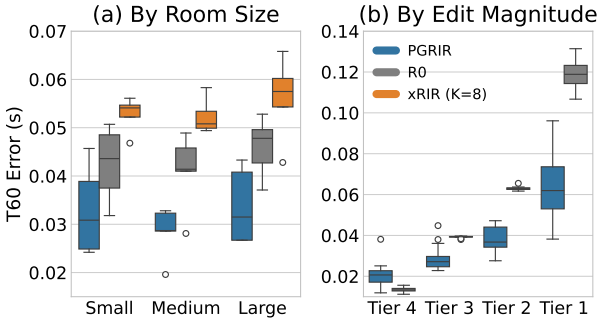

Figure 3: *Boxplots of* T_{60} *error under various conditions. (a) groups the results by room size, and (b) stratifies them by the amount of edit-induced acoustic deformation in terms of* T_{60}*.*

4.1. Results

PG-RIR vs RIR Interpolation-based prior art: As edit-conditioned RIR estimation is a new task lacking a well-established task-specific baseline, we compare PG-RIR against xRIR [25], a cross-scene RIR interpolation model. Specifically, xRIR predicts a target RIR from sparse reference RIRs observed in the same scene as the target. We retrain an audio-only xRIR variant without vision-relevant components on their dataset and fine-tune it on our RIR dataset.

From Table 1, we see that PG-RIR achieves the lowest error on all three metrics, outperforming both the identity baseline ($\mathbf{r}_0$) and the xRIR variants. Notably, xRIR lags behind identity baseline although it accesses reference RIRs for multiple source-receiver positions in the edited geometry G_1, unlike our approach. These results show that the pre-edit RIR at the same source-receiver position, $\mathbf{r}_0$, is an essential cue compared with sparse post-edit reference RIRs.

Figure 3 provides comprehensive analysis across (a) different room size and (b) different amounts of acoustic variation in terms of T_{60}. From the left panel, we observe that PG-RIR

Table 2: *DiffRIR baseline on two representative scenes. #(S) represents the number of surface/material parameters.*

#(S)	Method	$T_{60} \downarrow$	$C_{50} \downarrow$	EDT $\downarrow$
9	Identity ($\mathbf{r}_0$)	0.0261	0.826	0.022
	DiffRIR	0.0878	1.731	0.027
	PG-RIR	**0.0135**	**0.459**	**0.012**
34	Identity ($\mathbf{r}_0$)	0.0414	1.993	0.029
	DiffRIR	0.1826	7.160	0.092
	PG-RIR	**0.0315**	**1.615**	**0.021**

consistently outperforms the identity baseline and xRIR regardless of room sizes. In the right panel, we group $(\mathbf{r}_0, \mathbf{r}_1)$ pairs by the aforementioned tiers in Section 3. The performance gap between PG-RIR and the baseline widens for moderate and high-magnitude edits, suggesting that the model successfully learns to compensate for the edit-induced acoustic variation.

PG-RIR vs differentiable acoustic rendering: We additionally compare PG-RIR against DiffRIR [16], a differentiable acoustic rendering-based framework that can be adapted to geometry editing via a two-stage procedure. First, given the room geometry G_0 and the original RIR $\mathbf{r}_0$, DiffRIR performs *inverse rendering* to estimate surface material parameters by minimizing the reconstruction error of a differentiable acoustic ray tracer. The fitted parameters are then transferred to the edited geometry G_1, and a *forward rendering* pass synthesizes $\hat{\mathbf{r}}_1$.

One limitation of DiffRIR is that it requires a computationally expensive *per-scene* inverse rendering step, which becomes inaccurate and impractical as the number of surfaces and material parameters grows. Naturally this fitting does not scale to our entire, diverse, cross-scene benchmark, and hence we report DiffRIR on two representative cases: (i) a simplified *Classroom* scene (9 surfaces) aligned with its intended setup [16], and (ii) *Pomaria_2_int* from our benchmark (containing 34 surfaces), which is representative of realistic indoor complexity.

Table 2 shows that DiffRIR underperforms both the identity baseline and PG-RIR, even in the simplified 9-surface setting. The gap widens substantially on the 34-surface scene, showing that the difficulty of inverse-rendering with many surface/material parameters limits the performance of DiffRIR. PG-RIR consistently achieves the best performance even in the simple 9-surface scene, which is out of training data distribution (room meshes from the iGibson dataset).

Beyond quantitative metrics, Fig. 4 illustrates the learned transformation of PG-RIR for a representative high-magnitude scene edit. As the model computes a linear combination in the log-magnitude spectral domain, its (signed) attention weights function as power-law exponents in the linear amplitude domain. As shown, the network assigns negative weights to the prior observation $\mathbf{r}_0$ and positive weights to the target proxies. Mathematically, this operates as adaptive subtractive equalization; the model actively divides out (suppresses) the reverberant energy of $\mathbf{r}_0$ while multiplying in the new acoustic signature from the target proxies. Furthermore, this filtering has a non-uniform structure across frequency bins, confirming that PG-RIR treats edits as complex, frequency-dependent deformations rather than global gain adjustments.

4.2. Generalization to Translation Edits

Although trained exclusively on object removals, real-world scenes frequently undergo *translation* edits (e.g., sliding furniture). Since PG-RIR is conditioned by homogeneous proxies rather than strict geometric heuristics, it should naturally

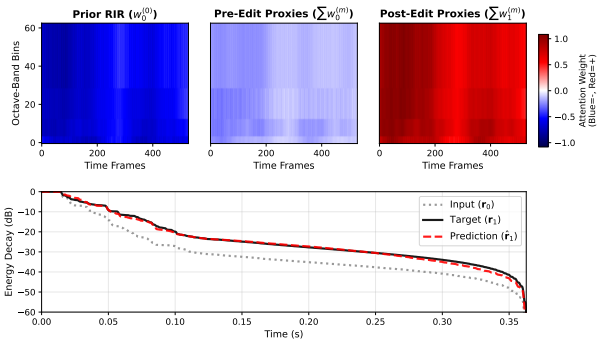

Figure 4: ***Top:*** *Temporal weights applied to the real RIR ($w_0^{(0)}$) and aggregated pre- and post-edit proxies ($\sum w_0^{(m)}$, $\sum w_1^{(m)}$). Red indicates positive mixing (energy addition); blue indicates negative mixing (spectral subtraction).* ***Bottom:*** *The resulting Energy Decay Curve (EDC) shows that PG-RIR successfully warps the pre-edit decay to match the physical target.*

Table 3: *3-step sequential edit chain ($G_0 \to T_1 \to T_2 \to G_0$) Metrics are macro-averaged across three unseen scenes. The Identity baseline ($\mathbf{r}_{t-1}$) represents the physical magnitude of the acoustic change induced by the edit, while PG-RIR ($\hat{\mathbf{r}}_t$) evaluates the model's ability to track that change using its own prediction from the previous step.*

Edit Step	Method	$T_{60} \downarrow$	$C_{50} \downarrow$	EDT $\downarrow$
Step 1 ($G_0 \to T_1$)	Identity ($\mathbf{r}_0$)	0.0755	4.214	0.056
	PG-RIR ($\hat{\mathbf{r}}_1$)	**0.0695**	**3.033**	**0.045**
Step 2 ($T_1 \to T_2$)	Identity ($\mathbf{r}_1$)	**0.0733**	5.251	0.069
	PG-RIR ($\hat{\mathbf{r}}_2$)	0.0747	**5.134**	**0.066**
Step 3 ($T_2 \to G_0$)	Identity ($\mathbf{r}_2$)	0.0746	5.721	0.067
	PG-RIR ($\hat{\mathbf{r}}_0$)	**0.0723**	**3.791**	**0.048**

generalize to unseen operations. We evaluate this using multi-step translation tuples ($G_0 \to T_1 \to T_2 \to G_0$) across three scenes (Beechwood_0_int, Merom_1_int, Wainscott_0_int), where T_1, T_2 are intermediate translated geometries. In this setup, the predicted RIR $\hat{\mathbf{r}}_t$ is fed back into the network as the input prior for the subsequent step $t+1$.

Table 3 tracks this 3-step chain. Step 1 demonstrates zero-shot generalization: despite the complete lack of translation samples in the training distribution, PG-RIR consistently outperforms the identity baseline across all three unseen scenes. Although PG-RIR follows geometric shifts across consecutive edits, repeated forward passes through the network inevitably accumulate spectral smoothing, which gradually smears early-energy transients.

5. Conclusions and Future Work

We introduce the novel task of edit-conditioned RIR estimation to update a measured RIR following geometry edits. Towards this end, we propose PG-RIR, which leverages simulated proxy RIRs from pre- and post-edit room meshes. Our experiments demonstrate that PG-RIR robustly predicts these acoustic updates across diverse room sizes and varying edit magnitudes. Additionally, the model demonstrates promising zero-shot generalization to unseen object translations and sequential edit chains. Future work will focus on improving performance by incorporating a diverse mixture of translation sequences via curriculum learning.

6. Generative AI Use Disclosure

The authors used generative AI to assist with language editing of the manuscript. All scientific content and conclusions were produced by the authors.

7. References

[1] H. Kuttruff, *Room Acoustics*. CRC Press, 2016.

[2] S. Koyama, E. De Sena, P. Samarasinghe, M. R. P. Thomas, and F. Antonacci, "Past, present, and future of spatial audio and room acoustics," in *Proc. IEEE ICASSP*, 2025.

[3] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[4] D. Botteldooren, "Finite-difference time-domain simulation of low-frequency room acoustic problems," *J. Acoust. Soc. Am.*, vol. 98, no. 6, pp. 3302–3308, 1995.

[5] L. L. Thompson, "A review of finite-element methods for time-harmonic acoustics," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1315–1330, 2006.

[6] G. Götz, D. G. Nielsen, S. Gudjónsson, and F. Pind, "Room-acoustic simulations as an alternative to measurements for audio-algorithm evaluation," *IEEE Access*, vol. 13, pp. 214 000–214 008, 2025.

[7] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. Van Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1929–1941, 2017.

[8] M. Jälmby, F. Elvander, and T. Van Waterschoot, "Low-rank room impulse response estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 957–969, 2023.

[9] S. Koyama, J. G. C. Ribeiro, T. Nakamura, N. Ueno, and M. Pezzoli, "Physics-informed machine learning for sound field estimation: Fundamentals, state of the art, and challenges," *IEEE Signal Process. Mag.*, vol. 41, no. 6, pp. 60–71, 2024.

[10] A. Richard, P. Dodds, and V. K. Ithapu, "Deep impulse responses: Estimating and parameterizing filters with deep networks," in *Proc. IEEE ICASSP*, 2022, pp. 3209–3213.

[11] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," *Proc. NeurIPS*, vol. 35, pp. 3165–3177, 2022.

[12] K. Su, M. Chen, and E. Shlizerman, "INRAS: Implicit neural representation for audio scenes," *Proc. NeurIPS*, vol. 35, pp. 8144–8158, 2022.

[13] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu, "AV-NeRF: Learning neural fields for real-world audio-visual scene synthesis," in *Proc. NeurIPS*, 2023.

[14] Z. Chen, I. D. Gebru, C. Richardt, A. Kumar, W. Laney, A. Owens, and A. Richard, "Real acoustic fields: An audio-visual room acoustics dataset and benchmark," in *Proc. CVPR*, 2024.

[15] Z. Lan, C. Zheng, Z. Zheng, and M. Zhao, "Acoustic volume rendering for neural impulse response fields," *Proc. NeurIPS*, vol. 37, pp. 44 600–44 623, 2024.

[16] M. L. Wang, R. Sawata, S. Clarke, R. Gao, S. Wu, and J. Wu, "Hearing anything anywhere," in *Proc. CVPR*, 2024, pp. 11 790–11 799.

[17] S. Majumder, C. Chen, Z. Al-Halah, and K. Grauman, "Few-shot audio-visual learning of environment acoustics," *Proc. NeurIPS*, vol. 35, pp. 2522–2536, 2022.

[18] A. Ratnarajah, S. Ghosh, S. Kumar, P. Chiniya, and D. Manocha, "AV-RIR: Audio-visual room impulse response estimation," in *Proc. CVPR*, 2024, pp. 27 164–27 175.

[19] C. Ick, G. Wichern, Y. Masuyama, F. G. Germain, and J. Le Roux, "Direction-aware neural acoustic fields for few-shot interpolation of ambisonic impulse responses," in *Proc. ISCA Interspeech*, 2025.

[20] M. Guski and M. Vorländer, "The influence of single scattering objects for room acoustic measurements," in *Proc. Annual German Conf. Acoust. (DAGA)*, 2015, pp. 334–335.

[21] G. Götz, S. J. Schlecht, and V. Pulkki, "A dataset of higher-order ambisonic room impulse responses and 3D models measured in a room with varying furniture," in *Proc. IEEE I3DA*, 2021.

[22] K. Prawda, "Sensitivity of room impulse responses in changing acoustic environment," in *Proc. IEEE ICASSP*, 2025.

[23] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "Bundlefusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, p. 1, 2017.

[24] G. Baruch, Z. Chen, A. Dehghan, T. Dimry, Y. Feigin, P. Fu, T. Gebauer, B. Joffe, D. Kurz, A. Schwartz, and E. Shulman, "ARKitscenes - a diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data," in *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[25] X. Liu, A. Kumar, P. Calamia, S. V. Amengual, C. Murdock, I. Ananthabhotla, P. Robinson, E. Shlizerman, V. K. Ithapu, and R. Gao, "Hearing anywhere in any environment," in *Proc. CVPR*, 2025, pp. 5732–5741.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[27] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *Proc. ICLR*, 2017.

[28] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, no. 3, pp. 409–412, 1965.

[29] B. Shen, F. Xia, C. Li, R. Martín-Martín, L. Fan, G. Wang, C. Pérez-D'Arpino, S. Buch, S. Srivastava, L. Tchapmi *et al.*, "iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes," in *Proc. IEEE/RSJ IROS*, 2021, pp. 7520–7527.

[30] Z. Lan, C. Zheng, Z. Zheng, and M. Zhao, "Acoustic volume rendering for neural impulse response fields," in *Proc. NeurIPS*, 2024.