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

Dual-Geometry Manifolds for Few-shot RIR Prediction

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

TR2026-139 September 29, 2026

Abstract

Few-shot room impulse response (RIR) prediction estimates target acoustics from sparse reference RIRs. Current models fuse references in a static Euclidean space, assuming a zero-curvature latent geometry for all temporal phases of the RIR. Using Gromov d-hyperbolicity, we show the acoustic manifold's geometry evolves from a hierarchical tree structure during early reflections to a flat, diffuse late reverberation tail. Motivated by this heterogeneity, we propose Janus-RIR, a temporally gated, dual-geometry aggregation model. It employs a hyperbolic branch for early reflections and a Euclidean branch for smooth statistical averaging of the late tail, governed by a context-aware dynamic gate. Experiments on AcousticRooms show Janus-RIR achieves state-of-the-art. Aligning latent geometry with the physics of acoustics improves speech clarity (C50) while reducing late reverberation (T60) error by 26% over Euclidean baselines, adapting to environment-specific mixing times.

Interspeech 2026

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

Dual-Geometry Manifolds for Few-shot RIR Prediction

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

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

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

Abstract

Few-shot room impulse response (RIR) prediction estimates target acoustics from sparse reference RIRs. Current models fuse references in a static Euclidean space, assuming a zero-curvature latent geometry for all temporal phases of the RIR. Using Gromov δ-hyperbolicity, we show the acoustic manifold's geometry evolves from a hierarchical tree structure during early reflections to a flat, diffuse late reverberation tail. Motivated by this heterogeneity, we propose `Janus-RIR`, a temporally gated, dual-geometry aggregation model. It employs a hyperbolic branch for early reflections and a Euclidean branch for smooth statistical averaging of the late tail, governed by a context-aware dynamic gate. Experiments on AcousticRooms show `Janus-RIR` achieves state-of-the-art. Aligning latent geometry with the physics of acoustics improves speech clarity (C_{50}) while reducing late reverberation (T_{60}) error by 26% over Euclidean baselines, adapting to environment-specific mixing times.

Index Terms: Room impulse response, non-Euclidean geometry, cross-scene generalization, spatial audio.

1. Introduction

A room impulse response (RIR) is a transfer function between a localized sound source and a receiver, characterizing how sound waves propagate, reflect, diffract, and eventually decay. In recent years, the challenge of cross-scene, few-shot RIR estimation, which aims to synthesize the acoustic response at an unseen target position using only a sparse set of reference RIR measurements from that same environment, has emerged as a critical setting for spatial audio and augmented reality tasks [1–3]. Recently, models such as xRIR [4] have established a strong baseline by fusing visual geometric context with acoustic reference features in a shared latent space.

Despite these advances, current attention-based aggregators implicitly map these reference RIRs into a fixed zero-curvature Euclidean space [5]. This architectural choice conflicts with a fundamental physical characteristic of room acoustics: RIRs are temporally heterogeneous. The early portion of an RIR is characterized by sparse, specular reflections following deterministic propagation paths. In classical geometrical acoustics, such as the image-source method [6], the early sound field is modeled via an exponentially growing hierarchy of virtual sources, expanding at a rate of $\mathcal{O}((S-1)^i)$ for S planar surfaces at reflection order i. Consequently, the underlying mathematical topology of these discrete specular paths is intrinsically tree-like. Conversely, as the reflection order increases, the acoustic field transitions at the *mixing time* into a dense reverberation tail. In this late regime, the high density of overlapping reflections breaks down the discrete hierarchy, resulting in a statistically uniform, diffuse sound field [7].

Because Euclidean geometry is mathematically incapable of embedding hierarchical trees without severe distortion [8,9], current models forcibly over-average the distinct early spatial paths. This geometric smearing [10] destructively interferes with early acoustic transients. Because speech clarity metrics like C_{50} are strictly defined by the ratio of preserved early reflection energy to late reverberation [11], this forced Euclidean averaging leads to catastrophic degradation in clarity prediction.

To formally quantify this structural evolution, we analyze the local acoustic manifold using Gromov δ-hyperbolicity [12]. As detailed in Section 2, we empirically prove that the geometry transitions from a highly hyperbolic (tree-like) state during early reflections to a flat (Euclidean) diffuse tail, with the temporal rate of this transition governed by the room's volume.

Motivated by this physical reality, we propose `Janus-RIR`, a temporally gated, mixed-curvature aggregation module. Instead of forcing all temporal phases into a single space, `Janus-RIR` introduces two parallel reference aggregation branches: (i) a **hyperbolic branch** (Poincaré ball) that preserves the discrete, branch-selective geometry of early reflections, and (ii) a **Euclidean branch** that enables smooth, uniform averaging suited for the diffuse late tail. Since the mixing time varies drastically across different environments,

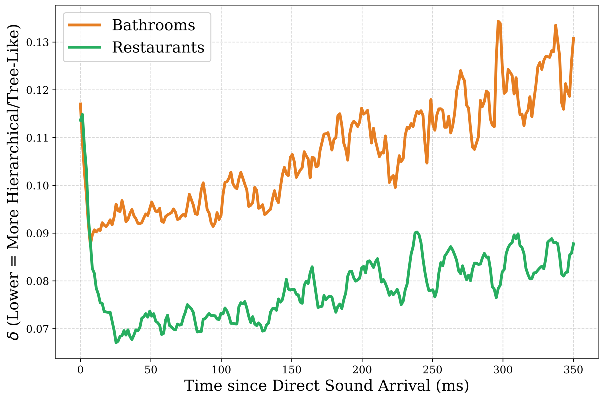

Figure 1: *Temporal transition of the local acoustic manifold ($\hat{\delta}$-hyperbolicity), computed from RIRs measured at 9 microphones. RIRs transition from a highly hyperbolic (tree-like, low $\hat{\delta}$) state during early reflections to a Euclidean (flat, high $\hat{\delta}$) diffuse tail. The rate of transition is strictly governed by the room's physical volume, with large environments (Restaurants) retaining their discrete tree structure significantly longer than small environments (Bathrooms).*

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

we introduce a **context-aware dynamic signature gate** ($\alpha(t)$) to temporally route features between the manifolds based on the room's (implicitly) inferred volume.

Extensive experiments on the AcousticRooms dataset [4] show that aligning the latent aggregation geometry with acoustic physics establishes a new state-of-the-art, reducing T_{60} error by 26% and drastically improving clarity (C_{50}) over competing Euclidean baselines. Furthermore, qualitative analysis of the learned latent space reveals that the model implicitly learns the physical mixing of its environment, actively holding the hyperbolic structure gate open longer in environments that physically support extended specular reflections.

2. Capturing Geometric Evolution of RIRs

In geometrical acoustics [6, 13], the early stage of sound propagation from a point source is modeled as an exponentially growing hierarchy of discrete specular reflection paths, mathematically structured as an image-source tree [14]. When a set of N microphones is distributed across a room, each microphone essentially samples a distinct, position-dependent subset of branches from this theoretical acoustic tree of virtual sources [15–17]. Because these specular reflections are highly directional, a specific sequence of bounces (e.g., source to floor to left wall to receiver) that is clearly audible at a specific microphone might be practically imperceptible or entirely occluded at another. We refer to the collective acoustic features captured by these microphones as the *local manifold*. During early reflections, the distances between the acoustic features across this manifold are governed by the strict, parent-child routing of the room's geometry. To formally validate that this discrete spatial routing manifests as a hierarchical topology in the audio features, we conduct a continuous δ-hyperbolicity analysis.

For a metric space (X, d), Gromov δ-hyperbolicity measures how "tree-like" it is by evaluating distance sums for any four points $x, y, z, w \in X$. Let the three sums of disjoint pairs (e.g., $d(x, y) + d(z, w)$) be sorted as: $S_{(1)} \geq S_{(2)} \geq S_{(3)}$. The global δ of the manifold is the maximum $\frac{1}{2}(S_{(1)} - S_{(2)})$ across all possible 4-tuples [12]. A space is highly hierarchical (hyperbolic) if δ approaches zero, whereas flat (Euclidean) spaces yield high δ values.

To analyze the temporal evolution of RIRs, we construct the local manifold X_t at each time step t using the features of the N spatially distributed microphones from the AcousticRooms dataset [4]. To isolate the physical reflection paths and prevent asynchronous Time-of-Flight (ToF) delays from artificially inflating geometric sparsity, we temporally shift and align all N signals following [4]. At each frame step, the features for the distance metric are defined as the flattened magnitude STFT frames over a short sliding window ($w = 5$). Finally, to ensure scale invariance as the acoustic energy naturally decays, we normalize the metric by the diameter of the manifold at that time frame: $\hat{\delta}(t) = \delta(t)/\mathrm{diam}(X_t)$.

Figure 1 plots the normalized $\hat{\delta}$-hyperbolicity across the smallest and largest room categories from the AcousticRooms dataset [4]. At the arrival of the direct sound ($t \approx 0$ ms), the initial wavefront acts as a single, uniform pulse across the array, resulting in a moderate starting $\hat{\delta}$ value. During the subsequent tree formation phase ($t \approx 20-100$ ms), early specular reflections arrive and the local manifold branches out. These sparse, highly position-dependent reflection paths render the data hierarchical, causing $\hat{\delta}(t)$ to drop sharply to a highly hyperbolic minimum. Finally, as the sound decays into a diffuse cloud ($t > 100$ ms), waves undergo higher-order boundary scattering and the room surpasses its mixing time [7]. Distinct reflection paths overlap, and the strict hierarchy breaks down into a dense, statistically uniform, Euclidean diffuse field [18], causing $\hat{\delta}(t)$ to steadily rise.

Crucially, this analysis demonstrates that the temporal rate of this transition depends on the room's physical profile. In smaller environments with fast diffusion, such as Bedrooms and Bathrooms, the reflections overlap rapidly, causing the geometry to evolve towards a flat Euclidean state early in the timeline. Conversely, in large environments like Restaurants, the extended distances allow the discrete, sparse specular branches to persist significantly longer, holding the local manifold in a highly hyperbolic state deep into the acoustic tail.

This demonstrates that static, zero-curvature latent geometries are fundamentally mismatched to the physics of RIRs. A purely Euclidean space fails to preserve the hierarchical constraints of early reflections, while a purely hyperbolic space over-penalizes the statistical averaging required for the late diffuse tail. In geometric representation learning, it is well-established that embedding data with heterogeneous topologies, containing both tree-like and flat structures requires models that leverage multiple curvature spaces [19]. Drawing inspiration from this, and because the acoustic transition is a continuous temporal shift governed by the room's specific volume and boundary absorption, we propose modeling RIRs using a **temporally gated, dual-geometry architecture**. By late-fusing representations aggregated in parallel hyperbolic and Euclidean spaces, our approach enables continuous geometric adaptation along the same temporal axis.

3. Methodology

Motivated by the continuous geometric regime shift, we propose `Janus-RIR`, an architecture that dynamically adapts its latent geometry to match the temporal physics of room acoustics. Given K reference RIRs and an unseen target configuration, our objective is to predict the target RIR's log-magnitude spectrogram, $\hat{\mathbf{Y}} \in \mathbb{R}^{F \times T}$, where F denotes the number of frequency bins and T the number of frames.

To isolate our geometric aggregation, we adopt the feature backbone from xRIR [4]. We align the K reference RIRs using Time-of-Flight delays relative to the target. These shifted references are converted to log-magnitude spectrograms $\mathbf{S}_k \in \mathbb{R}^{F \times T}$ and processed via a ResNet-18 [20] encoder. The acoustic features are concatenated with MLP-encoded coordinate embeddings to produce reference embeddings $\mathbf{h}_k \in \mathbb{R}^C$. We derive a target query embedding $\mathbf{h}_0 \in \mathbb{R}^C$ from the target coordinates, and a learned temporal basis $\boldsymbol{\tau}(t) \in \mathbb{R}^C$ spanning T frames. These features are routed to parallel Hyperbolic and Euclidean branches.

Hyperbolic Branch: To preserve the hierarchical geometry of early reflections and prevent the smearing of distinct propagation paths [6, 8], we map the Euclidean features to the Poincaré ball ($\mathbb{D}$) via the exponential map [21] at the origin: $\bar{\mathbf{h}}_k = \exp_{\mathbf{0}}(\mathbf{h}_k)$ and $\bar{\mathbf{h}}_0 = \exp_{\mathbf{0}}(\mathbf{h}_0)$. We compute hyperbolic attention logits, combining the scaled negative hyperbolic distance ($d_{\mathbb{D}}$) with a temporal alignment bias:

$$a_H(k,t) = -\beta\, d_{\mathbb{D}}\big(\bar{\mathbf{h}}_0, \bar{\mathbf{h}}_k\big) + \frac{1}{C}\big\langle \mathbf{h}_k, \boldsymbol{\tau}(t)\big\rangle \quad (1)$$

$$W_H(k,t) = \mathrm{softmax}_k\big(a_H(k,t)\big) \quad (2)$$

where β is a learnable scale. To aggregate reference spectrogram vectors in a curvature-consistent way, we stabilize the log-magnitude values using a learnable scale γ and a fixed shift b to keep points safely within the manifold boundary. We map these stabilized features to the Poincaré ball [22], $\bar{\mathbf{s}}_k(t) = \exp_{\mathbf{0}}\big(\gamma(\mathbf{s}_k(t)+b)\big)$, compute the weighted Einstein midpoint (Möbius gyromidpoint) [23], and project it back to the tangent space:

$$\mathbf{y}_H(t) = \gamma^{-1}\log_{\mathbf{0}}\Big(\mathrm{Midpoint}_{\mathbb{D}}\Big(\{\bar{\mathbf{s}}_k(t)\}_{k=1}^{K}, \{W_H(k,t)\}_{k=1}^{K}\Big)\Big) - b \tag{3}$$

Euclidean Branch: To enable the smooth statistical averaging required for the dense, uniform energy cloud of late reverberation [11, 24, 25], we compute Euclidean attention logits using standard dot-product similarity, and obtain the aggregated frame via a weighted sum:

$$\begin{aligned} a_E(k,t) &= \frac{1}{C}\langle \mathbf{h}_k, \mathbf{h}_0\rangle + \frac{1}{C}\langle \mathbf{h}_k, \boldsymbol{\tau}(t)\rangle, \\ W_E(k,t) &= \mathrm{softmax}_k\big(a_E(k,t)\big) \end{aligned} \tag{4}$$

The Euclidean aggregated spectrogram frame is obtained via a standard weighted sum:

$$\mathbf{y}_E(t) = \sum_{k=1}^{K} W_E(k,t)\,\mathbf{s}_k(t) \tag{5}$$

Context-Aware Dynamic Signature Gate: Because the physical mixing time varies drastically across different scenes [7, 26], we dynamically fuse our parallel branches using a Context-Aware Signature Gate $\mathcal{G}_\phi$. This network predicts a mixture coefficient $\alpha(t) \in [0,1]$ conditioned on the temporal embedding, the target query, and an acoustic context vector obtained by max-pooling the reference features:

$$\alpha(t) = \sigma\Big(\mathcal{G}_\phi\big([\boldsymbol{\tau}(t) \parallel \mathbf{h}_0 \parallel \max_k \mathbf{h}_k]\big)\Big) \tag{6}$$

where $\parallel$ denotes concatenation. Building on the framework of mixed-curvature product manifolds [19, 27], we employ a *product fusion* that concatenates our dynamically gated expert outputs and learns a projection $\mathbf{W}_p$ back to $\mathbb{R}^F$:

$$\begin{aligned} \tilde{\mathbf{y}}(t) &= \big[\alpha(t)\mathbf{y}_H(t) \parallel (1-\alpha(t))\mathbf{y}_E(t)\big] \in \mathbb{R}^{2F}, \\ \hat{\mathbf{y}}(t) &= \mathbf{W}_p\tilde{\mathbf{y}}(t) + \mathbf{b}_p \end{aligned} \tag{7}$$

Finally, to prevent the hyperbolic expert from over-averaging distinct branches during the early regime, we penalize flat attention distributions. We penalize the entropy of W_H using a time-decaying mask $m(t)$ that drops linearly from 1.0 at $t = 0$ to 0.1 at $t = T$:

$$\mathcal{L}_{ent} = \frac{\lambda}{T}\sum_{t=1}^{T} m(t)\left(-\sum_{k=1}^{K} W_H(k,t)\log\big(W_H(k,t)+\epsilon\big)\right) \tag{8}$$

The network is trained end-to-end using $\mathcal{L}_{ent}$ alongside standard $\mathcal{L}_1$ spectrogram reconstruction and energy decay losses [4]. During inference, time-domain RIR waveforms are recovered from the predicted magnitude spectrograms via the Griffin-Lim [28] reconstruction algorithm.

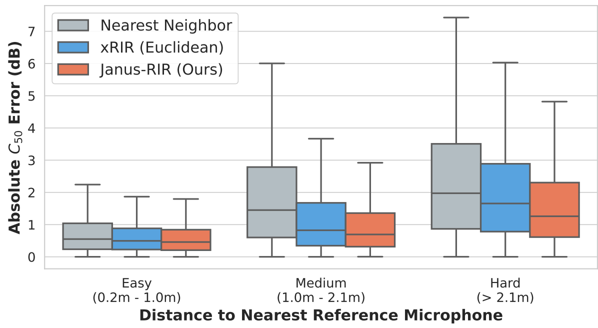

Figure 2: *Absolute C_{50} error distribution across spatial extrapolation difficulty tiers at $K = 8$.*

4. Experiments and Results

We evaluate cross-scene interpolation on the AcousticRooms dataset [4] (260 rooms, 10 categories) using $K \in \{1, 4, 8\}$ reference shots. We report errors in Early Decay Time (EDT, seconds), Clarity (C_{50}, dB), and Reverberation Time (T_{60}, percentage). Our code is available online[1].

Baselines: We compare against Few-shot RIR [1] (Euclidean cross-attention) and xRIR [4] (the current state-of-the-art, which fuses panoramic depth maps and acoustic features into a single Euclidean space). To isolate our geometric contributions, we evaluate `Janus-RIR` (incorporating xRIR's ViT [29] encoded depth maps) and `Janus-RIR` (NoVision) (relying solely on 3D coordinates and the reference RIRs).

As shown in Table 1, both variants of `Janus-RIR` establish a new state-of-the-art across all metrics and reference-shot configurations. The benefits of modeling the continuous geometric evolution of the RIR become particularly evident as the number of available references (K) increases. At $K = 8$, the `Janus-RIR` (NoVision) model achieves a T_{60} error of 7.75%, a remarkable 26% relative error reduction compared to the Euclidean xRIR baseline (10.53%). Crucially, this improvement is coupled with significant reductions in Early Decay Time error (0.055 s $\rightarrow$ 0.045 s) and Clarity error (1.457 dB $\rightarrow$ 1.126 dB). Because C_{50} explicitly measures the ratio of early reflection energy to late tail energy [11], this simultaneous improvement proves that our hyperbolic branch successfully preserves the strict structural hierarchy of early transients, while the dynamic Euclidean gate smoothly transitions to the necessary diffuse averaging for the late tail.

Furthermore, Fig. 2 breaks down the absolute C_{50} error across strict spatial difficulty tiers, defined by the distance from the target position to the nearest available reference microphone. We observe that while the Euclidean baseline, xRIR, degrades severely and exhibits massive error variance when synthesizing targets farther than 2.1m from a reference (the 'Hard' tier), `Janus-RIR` maintains significantly tighter error bounds.

4.1. Qualitative Analysis

To evaluate geometric adaptation, we aggregate internal latent trajectories across all unseen test samples from the smallest (Bathrooms, avg. T_{60} = 0.41s) and largest (Restaurants, avg. T_{60} = 0.73s) environments (Fig. 3). The context-aware gate α (left) acts as a general temporal prior, smoothly decaying from early transients to late reverberation. However, true geometric adaptation occurs within the hyperbolic manifold (middle). By

[1] `https://github.com/merlresearch/janus-rir`

Table 1: *Quantitative comparison of few-shot RIR synthesis across $K \in \{1,4,8\}$ references. Our dynamically gated, dual-geometry models (`Janus-RIR`) consistently outperform Euclidean baselines across all acoustic metrics, with the NoVision variant achieving state-of-the-art performance.*

Model	**K = 1**			**K = 4**			**K = 8**		
	EDT (s) ↓	**C50 (dB)** ↓	**T60 (%)** ↓	**EDT (s)** ↓	**C50 (dB)** ↓	**T60 (%)** ↓	**EDT (s)** ↓	**C50 (dB)** ↓	**T60 (%)** ↓
Few-shot RIR [1]	0.130	3.220	20.10	0.136	3.560	19.30	0.187	4.470	21.15
xRIR [4]	0.075	1.840	13.47	0.068	1.330	13.28	0.055	1.450	10.53
xRIR (NoVision) [4]	-	-	-	-	-	-	0.050	1.490	11.93
`Janus-RIR`	0.054	**1.53**	12.60	**0.048**	**1.250**	8.83	0.047	1.200	7.97
`Janus-RIR` (NoVision)	**0.054**	1.590	**11.28**	0.049	1.280	**8.78**	**0.045**	**1.120**	**7.75**

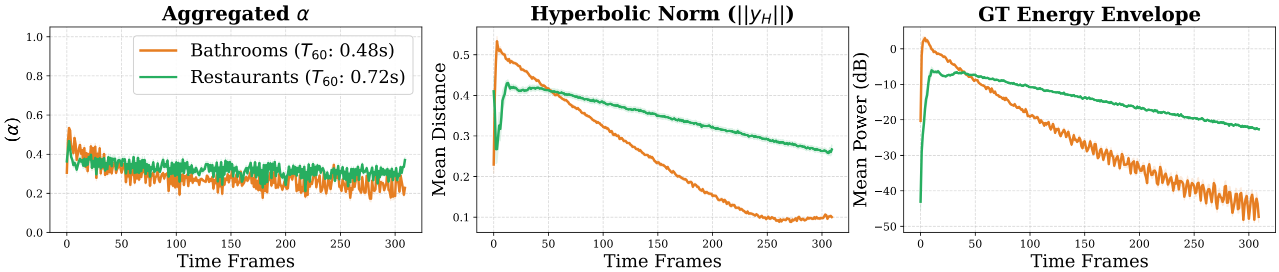

Figure 3: *Aggregated latent behaviors of `Janus-RIR` across unseen test samples from Bathrooms (small) and Restaurants (large). While the α gate (left) acts as a general temporal prior, true geometric adaptation occurs within the Hyperbolic branch (middle). `Janus-RIR` collapses structural certainty ($||y_H||$) for fast-diffusing small rooms but actively sustains it for large rooms, mirroring the ground-truth energy decay (right) and our empirical δ-hyperbolicity analysis (Fig. 1).*

Table 2: *Ablation studies on `Janus-RIR` for $K = 8$.*

Configuration	**EDT (s)** ↓	**C50 (dB)** ↓	**T60 (%)** ↓
`Janus-RIR` (NoVision)	**0.045**	**1.127**	**7.75**
(A) Hyperbolic Only ($\alpha = 1$)	0.059	1.515	10.52
(B) Dual-Euclidean Branches	0.053	1.328	8.52
(C) Time-Only Gate	0.057	1.379	9.46

tracking the structural certainty (mean vector norm $||y_H||$) [30], we observe that for smaller rooms, the hyperbolic distance collapses rapidly toward the origin, allowing the Euclidean branch to dominate as sound quickly diffuses. Conversely, for larger rooms (Restaurant category), the network actively holds the representation far from the origin, preserving the sparse tree structure deep into the temporal tail. This mirrors the ground-truth energy decay (right) and our empirical δ-hyperbolicity analysis (Fig. 1). Without explicit volume supervision, `Janus-RIR` implicitly learns the physical mixing times, collapsing its tree structure for fast-diffusing rooms while sustaining it for environments with extended specular reflections.

4.2. Ablation Study

To validate the specific architectural choices of our mixed-curvature formulation, we conduct ablation studies on the top-performing **`Janus-RIR` (NoVision)** configuration at $K = 8$. As seen in Table 2, removing any component of our temporal geometric alignment degrades performance.

The Failure of Static Curvature (Hyperbolic Only): In Ablation (A), we override the dynamic signature gate and force $\alpha(t) = 1.0$ for all time steps, effectively mapping the entire RIR sequence into a purely hyperbolic (Poincaré) space. This results in a massive degradation across all metrics, with the T_{60} error degrading from 7.75% back up to 10.52%, mirroring the failure of the **Euclidean Only** baseline, xRIR. We thus validate that while hyperbolic space perfectly isolates early hierarchical reflections, it actively prevents the statistical averaging required to accurately synthesize the diffuse late tail.

Geometric Alignment vs. Parameter Count: To ensure our performance gains stem from geometric alignment rather than a mere increase in network capacity or an ensemble effect, Ablation (B) replaces the hyperbolic branch with a second Euclidean branch. Performance drops significantly, particularly in clarity (C_{50} worsens from 1.127 dB to 1.328 dB). This confirms that Euclidean space, regardless of parameter count or parallel routing, inherently smears early specular reflections. Both (A) and (B) validate the effectiveness of our dual-geometry model.

The Necessity of Acoustic Context: In Ablation (C), we strip the acoustic context features from the gating network $\mathcal{G}_\phi$, forcing $\alpha(t)$ to rely exclusively on the global temporal embedding. The resulting degradation (EDT 0.045 s $\to$ 0.057 s, T_{60} 7.75% $\to$ 9.46%) demonstrates that the transition from discrete reflections to diffuse scattering does not happen at a globally fixed timestamp. Rather, the mixing time depends heavily on the physical volume and boundaries of the specific room, necessitating the global acoustic context of the reference set.

5. Conclusions and Future Work

In this work, we demonstrated that mapping temporally heterogeneous room impulse responses into a static Euclidean latent space fundamentally smears the hierarchical structure of early acoustic reflections. To address this, we introduced `Janus-RIR`, a dual-geometry architecture that dynamically routes acoustic features between hyperbolic and Euclidean manifolds. Our approach achieves state-of-the-art few-shot synthesis, significantly reducing both early speech clarity (C_{50}) and late reverberation (T_{60}) errors. Moreover, aggregation of the learned latent spaces shows that our network actively adapts its hyperbolic structural certainty to match the specific acoustic volume of the room. Future work will consider time-domain interpolation schemes for the early reflections, for example via optimal transport, inspired by [31].

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] 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.

[2] 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.

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

[4] 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.

[5] J. M. Lee, *Riemannian manifolds: An introduction to curvature*. Springer Science & Business Media, 2006.

[6] 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.

[7] J.-D. Polack, "Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics," *Applied Acoustics*, vol. 38, no. 2-4, pp. 235–244, 1993.

[8] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," *Proc. NeurIPS*, vol. 30, 2017.

[9] A. Tifrea, G. Bécigneul, and O.-E. Ganea, "Poincaré glove: Hyperbolic word embeddings," *Proc. ICLR*, 2019.

[10] F. Sala, C. De Sa, A. Gu, and C. Ré, "Representation tradeoffs for hyperbolic embeddings," in *Proc. ICML*. PMLR, 2018, pp. 4460–4469.

[11] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[12] M. Gromov, "Hyperbolic groups," in *Essays in group theory*. Springer, 1987.

[13] M. Vorländer, *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media, 2008.

[14] P. Charalampous and D. Michael, "Representation of the tree structure of an image source algorithm," in *11th Sound and Music Computing Conference*, 2014.

[15] F. Antonacci, M. R. Thomas, E. A. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Geometrically constrained room modeling with compact microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1449–1460, 2012.

[16] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12 186–12 191, 2013.

[17] D. Marković, F. Antonacci, A. Sarti, and S. Tubaro, "3D beam tracing based on visibility lookup for interactive acoustic modeling," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 10, pp. 2262–2274, 2016.

[18] M. R. Schroeder, "Frequency-correlation functions of frequency responses in rooms," *J. Acoust. Soc. Am.*, vol. 34, no. 12, pp. 1819–1823, 1962.

[19] A. Gu, F. Sala, B. Gunel, and C. Ré, "Learning mixed-curvature representations in product spaces," in *Proc. ICLR*, 2018.

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

[21] M. Nickel and D. Kiela, "Learning continuous hierarchies in the Lorentz model of hyperbolic geometry," in *Proc. ICML*. PMLR, 2018, pp. 3779–3788.

[22] K. Desai, M. Nickel, T. Rajpurohit, J. Johnson, and S. R. Vedantam, "Hyperbolic image-text representations," in *Proc. ICML*. PMLR, 2023, pp. 7694–7731.

[23] C. Gulcehre, M. Denil, M. Malinowski, A. Razavi, R. Pascanu, K. M. Hermann, P. Battaglia, V. Bapst, D. Raposo, A. Santoro *et al.*, "Hyperbolic attention networks," *Proc. ICLR*, 2019.

[24] C.-H. Jeong, "Diffuse sound field: challenges and misconceptions," in *45th International Congress and Exposition on Noise Control Engineering*. German Acoustical Society (DEGA), 2016, pp. 1015–1021.

[25] M. Nolan, M. Berzborn, and E. Fernandez-Grande, "Isotropy in decaying reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 148, no. 2, pp. 1077–1088, 2020.

[26] P. Götz, K. Kowalczyk, A. Silzle, and E. A. Habets, "Mixing time prediction using spherical microphone arrays," *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. EL206–EL212, 2015.

[27] H. Sáez de Ocáriz Borde, A. Arroyo, I. Morales, I. Posner, and X. Dong, "Neural latent geometry search: Product manifold inference via Gromov-Hausdorff-informed Bayesian optimization," *Proc. NeurIPS*, vol. 36, pp. 38 370–38 403, 2023.

[28] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *Proc. ICLR*, 2021.

[30] O. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic entailment cones for learning hierarchical embeddings," in *Proc. ICML*. PMLR, 2018, pp. 1646–1655.

[31] A. Geldert, N. Meyer-Kahlen, and S. J. Schlecht, "Interpolation of spatial room impulse responses using partial optimal transport," in *Proc. IEEE ICASSP*, 2023.