HEADING-EQUIVARIANT ATTENTION FOR YAW-INVARIANT FEW-SHOT ROOM-ACOUSTIC SYNTHESIS

Anonymous authors
Paper under double-blind review

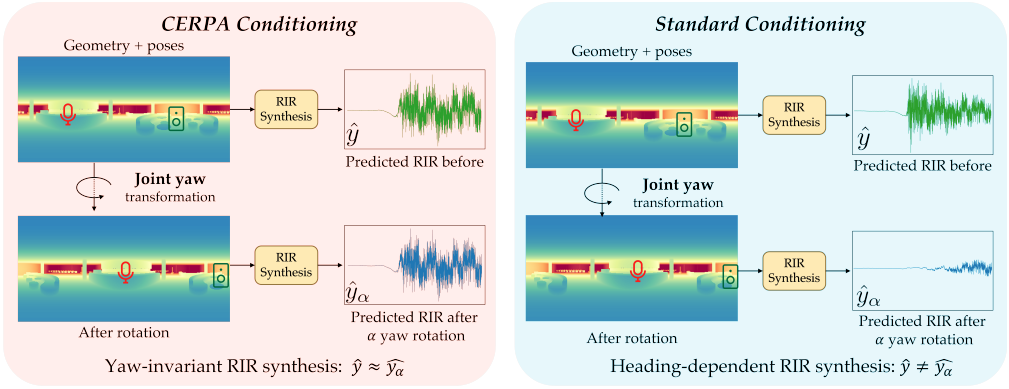

Figure 1: **Heading-equivariant conditioning for yaw-invariant acoustic synthesis.** Jointly rotating panorama and source–receiver poses in yaw changes only the coordinate frame, leaving the underlying acoustics unchanged; the predicted RIR should remain invariant. CERPA maintains nearly identical predictions (left), whereas standard conditioning violates this invariance (right).

ABSTRACT

State-of-the-art cross-room few-shot acoustic synthesis methods generate room impulse responses (RIRs) from sparse acoustic context and geometry conditioningf. However, their geometry conditioning often fails to respect the heading invariance of monaural acoustics: jointly applying a yaw rotation to the panoramic depth image and source–receiver poses can change the predicted RIRs. Such a change of heading reference changes only the coordinate frame, leaving the underlying acoustic characteristics unchanged; the conditional probabilistic RIR distribution or deterministic RIR output should therefore remain invariant. To address such physical inconsistency, we introduce CERPA (Cyclic-Equivariant Relative Positional Attention), a unified attention framework that explicitly builds this rotational invariance into the visual representation. CERPA modifies the attention mechanism of the geometry-conditioning ViT to produce cyclic rotation equivariant tokens, followed by an invariant readout for heading-invariant visual conditioning. When integrated into both deterministic and state-of-the-art probabilistic acoustic models, CERPA ensures yaw invariant acoustic synthesis. In both simulated and real-world few-shot RIR synthesis, our approach substantially improves heading consistency and outperforms non-equivariant baselines on key acoustic metrics.

1 INTRODUCTION

Imagine listening to the same piece of music in a concert hall and a recording studio. Although the source audio is unchanged, each space imparts its own acoustic character, creating a distinct listening experience. Reproducing this interaction between sound and space is essential for immersive virtual environments (Lentz et al., 2007; Thery & Katz, 2019), where realistic immersion depends on both visual and spatial acoustic information. Such acoustic properties are characterized by room impulse responses (RIRs) (Vorländer, 2020), which model how sound propagates from a source to a receiver.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Accurately modeling these responses (Liu et al., 2025; Brunetto, 2026) requires capturing how room geometry, surface materials, and source–receiver configurations jointly shape sound propagation.

Neural acoustic fields (Luo et al., 2022; Su et al., 2022; Liang et al., 2023b; Brunetto et al., 2025) address this challenge by learning spatially continuous representations of room acoustics, enabling RIR prediction at unobserved locations within a scene. Related audio-visual approaches (Liang et al., 2023a; Bhosale et al., 2024; Chen & Shlizerman, 2024) incorporate visual scene representations to synthesize spatial audio at novel viewpoints. However, many of these methods require densely sampled RIR recordings and separate training for each room, making them costly to scale across environments. This motivates models that can synthesize RIRs in new rooms from only a few acoustic observations, without additional training.

Recent work (Majumder et al., 2022; Huang et al., 2025) has explored this area to generalize trained acoustic model to new environment with sparse acoustic context, known as few-shot acoustic synthesis. Within this few-shot paradigm, methods such as xRIR (Liu et al., 2025) and FLAC (Brunetto, 2026) predict RIRs in novel environments using geometric conditioning and a few acoustic context, without scene-specific retraining. However, prior work has shown that the Vision Transformer (ViT) used for geometry conditioning can be sensitive to input shifts (Rojas-Gomez et al., 2024; Ding et al., 2023) and global rotations (Feng et al., 2025). Consistent with these studies, Figure 7 and our pilot study (Appendix C.1) show that changing only the global heading reference can substantially alter the predicted RIR and degrade synthesis quality. This sensitivity is unexpected, as acoustic predictions should remain invariant to the choice of heading reference. In fact, jointly changing the heading of the panorama and source–receiver coordinates leaves the physical acoustic configuration unchanged (Figure 1), so the conditional distribution of omnidirectional monaural RIRs should remain invariant. This symmetry provides a natural inductive bias for geometry conditioning: panoramic features should transform equivariantly under yaw-induced cyclic shifts, while acoustic predictions remain invariant. Standard visual encoders do not explicitly enforce this relationship.

To address this gap, we develop **CERPA** (**C**yclic-**E**quivariant **R**elative **P**ositional **A**ttention), a unified framework for heading-equivariant panorama transformers. We formulate patch-aligned yaw rotations as the finite cyclic subgroup $C_N \subset \mathrm{SO}(2)$, consisting of rotations in increments of $2\pi/N$, where N is the number of horizontal patch columns. We design attention conditions that ensure equivariant geometry conditioning and invariant output token readout under this group action. Unlike non-equivariant vision transformers, CERPA preserves the cyclic equivariance of geometry conditioning through token processing. This makes C_N-equivariance an architectural property maintained throughout learning. We then combine invariant readouts with heading-invariant pose conditioning to enable heading-invariant acoustic synthesis. To validate our framework, we evaluate two realizations (Additive and Rotary CERPA) on the AcousticRooms dataset. Both approaches guarantee C_N-invariant RIR predictions and outperform non-equivariant baselines on key acoustic metrics. To summarize, our main contributions are as follows:

- We propose CERPA, a unified framework for heading-equivariant attention under the discrete cyclic group C_N.
- We derive circular relative-bias and cylindrical rotary realizations, instantiated as CYLINDRICALVIT (Additive CERPA) and CYLDINO (Rotary CERPA), respectively.
- We demonstrate patch-aligned heading-invariant RIR synthesis in xRIR and FLAC on both simulation and real-world dataset with quality comparable to or better than non-equivariant counterparts.

2 RELATED WORK

Equivariant networks and relative positional encoding. Group convolutions, steerable networks, and equivariant self-attention encode known symmetries through architectural constraints (Cohen & Welling, 2016; Weiler & Cesa, 2019; Romero & Cordonnier, 2021; Hutchinson et al., 2021). Relative positional interactions can be implemented through additive terms (Shaw et al., 2018; Liu et al., 2021; Press et al., 2022) or rotary embeddings (Su et al., 2024; Heo et al., 2024; Siméoni et al., 2025), with extensions to learned Lie-group operators and cylindrical domains (Ostmeier et al., 2025; Weng et al., 2025). At a more general level, Algebraic Positional Encodings explicitly support finite cyclic domains (Kogkalidis et al., 2024), while GRAPE provides a unified view of additive and multiplicative positional formulations (Zhang et al., 2026). Beyond equivariant feature construction, downstream invariant representations can be obtained through symmetric aggregation

(Zaheer et al., 2017). While existing mechanisms focus on representation learning, they fail to connect cyclic-equivariant interactions with invariant conditioning for panorama-based physical predictions. CERPA bridges this gap through unified attention with additive and rotary realizations, followed by invariant readout to guarantee heading-invariant acoustic conditioning.

Panoramic vision and shift-equivariant transformers. Spherical networks construct rotation-equivariant representations (Cohen et al., 2018; Esteves et al., 2018), while panorama-specific architectures address projection distortion and boundary structure (Coors et al., 2018; Shen et al., 2022; Ling et al., 2023). These periodic boundary effects also motivate transformer-specific designs that restore shift equivariance (Ding et al., 2023; Rojas-Gomez et al., 2024) or introduce topology-aware rotary encodings for panorama generation (Kang et al., 2026; Hirschorn et al., 2026). These methods primarily address the topology, projection, or shift structure of panoramic images, but do not explicitly model the joint change of heading reference across panoramic geometry and associated source–receiver coordinates. CERPA resolves this limitation by explicitly modeling this coordinate-frame symmetry, coupling C_N-equivariant panoramic representations with heading-invariant geometric and pose conditioning.

Neural room acoustics and few-shot synthesis. Neural acoustic fields model spatial audio responses (Luo et al., 2022; Su et al., 2022; Liang et al., 2023b; Brunetto et al., 2025; Lan et al., 2024) with implicit neural field, while audio-visual approaches further incorporate scene geometry and appearance for spatial audio rendering (Liang et al., 2023a; Bhosale et al., 2024; Chen & Shlizerman, 2024). However, these methods are typically scene-specific, motivating few-shot approaches that generalize across rooms from sparse observations (Majumder et al., 2022; Huang et al., 2025). Recent examples include deterministic prediction in xRIR (Liu et al., 2025) and conditional flow-based generation in FLAC (Brunetto, 2026). Although xRIR and FLAC explicitly condition on panoramic geometry and source–receiver poses, their geometry encoders do not architecturally enforce consistency under a change of heading reference. CERPA addresses this gap by introducing heading-equivariant geometry representations and heading-invariant conditioning into both systems, while evaluating heading consistency alongside acoustic synthesis quality.

3 CERPA: Heading-Equivariant Attention

3.1 Problem Setup and Heading Symmetry

Few-shot acoustic synthesis. We consider few-shot acoustic synthesis conditioned on panoramic depth and a few reference recordings. Let $x_s, x_r \in \mathbb{R}^3$ denote the source and receiver positions, respectively, and let $q = (x_s, x_r)$ denote the query. The target $h \in \mathbb{R}^{L_h}$ is a monaural, omnidirectional RIR. The synthesis model is conditioned on the sparse scene context $\mathcal{O} = (\mathcal{A}, \mathcal{P}, G_{x_r})$, where $\mathcal{A} = \{h_1, \dots, h_{K_{\text{ctx}}}\}$ denotes the reference RIRs and $\mathcal{P} = \{x_s^{(1)}, \dots, x_s^{(K_{\text{ctx}})}\}$ their source positions (Brunetto, 2026). Each h_j is measured at the same receiver x_r from $x_s^{(j)}$, and G_{x_r} is a panoramic depth map registered to the receiver coordinate frame. We denote the model's conditional RIR distribution by $p_\theta(h \mid q, \mathcal{O})$ and study its consistency under changes of heading reference.

Heading symmetry. Jointly re-expressing the panorama and all associated source–receiver coordinates under a global yaw rotation leaves the physical acoustic configuration unchanged, so the predicted acoustics should remain invariant. Let $\mathcal{G}_{\text{yaw}} = \{g_k \mid k \in \mathbb{Z}_N\}, g_k \circ g_\ell = g_{[k+\ell]_N}$, where g_k denotes a yaw rotation by $\alpha_k = 2\pi k/N$ and $N = W/p_w$ is the number of horizontal patch columns. As such, $\mathcal{G}_{\text{yaw}}$ is isomorphic to the finite cyclic group C_N, i.e., $\mathcal{G}_{\text{yaw}} \cong C_N$. For $g_k \in \mathcal{G}_{\text{yaw}}$, let R_{g_k} denote its action on the conditioning input: it applies the corresponding yaw rotation to all spatial coordinates and the matching cyclic shift to G_{x_r}, while leaving the reference RIRs and their associations unchanged. Accordingly, we require

$$p_\theta(h \mid q, \mathcal{O}) = p_\theta(h \mid R_{g_k} q, R_{g_k} \mathcal{O}), \qquad g_k \in \mathcal{G}_{\text{yaw}}. \tag{1}$$

CERPA enforces this symmetry through cyclic-equivariant geometry encoding, invariant readout, and heading-invariant pose conditioning (Section 3.4).

3.2 CERPA: Cyclic-Equivariant Attention

Cyclic positional structure. Let $G_{x_r} \in \mathbb{R}^{H \times W}$ be the panoramic depth map defined in Section 3.1. For horizontal and vertical patch sizes p_w and p_h, define $N = \frac{W}{p_w}, H_t = \frac{H}{p_h}$, and let $[H_t] =$

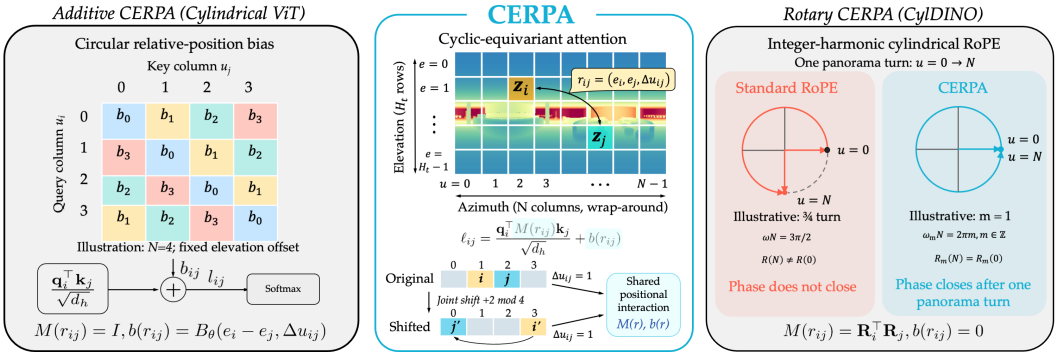

Figure 2: **Unified cyclic-equivariant attention for panoramic geometry.** CERPA preserves relative positional interactions under cyclic panorama shifts (center). Additive CERPA uses circular relative-position biases (left), while Rotary CERPA uses integer-harmonic cylindrical RoPE whose phases close after one panorama turn (right).

$\{0, \dots, H_t - 1\}$. The token grid is $\Omega = [H_t] \times \mathbb{Z}_N$, where $\mathbb{Z}_N = \{0, \dots, N-1\}$ indexes the azimuthal token columns modulo N. Each position $(e, u) \in \Omega$ is associated with a token feature $\mathbf{z}_{e,u} \in \mathbb{R}^d$. We denote the resulting feature map by $\mathbf{Z} : \Omega \to \mathbb{R}^d$. For $g_k \in \mathcal{G}_{\text{yaw}}$, the induced action on token positions is

$$g_k \cdot (e, u) = (e, [u + k]_N),$$

where $[\cdot]_N$ denotes reduction modulo N. This action on Ω induces a permutation operator P_k on token feature maps:

$$(P_k \mathbf{Z})(e, u) = \mathbf{Z}\big(g_k^{-1} \cdot (e, u)\big) = \mathbf{Z}(e, [u - k]_N)\,. \tag{2}$$

For tokens i and j at positions $p_i = (e_i, u_i)$ and $p_j = (e_j, u_j)$, their relative azimuth $\Delta u_{ij} = [u_j - u_i]_N$ is invariant under the common action of any $g_k \in \mathcal{G}_{\text{yaw}}$. Therefore, we define the heading-invariant pairwise positional descriptor

$$r_{ij} := (e_i, e_j, \Delta u_{ij})\,, \qquad \Delta u_{ij} = [u_j - u_i]_N. \tag{3}$$

Indeed, $[(u_j + k) - (u_i + k)]_N = [u_j - u_i]_N$, while e_i and e_j are unchanged because heading transformations act only on the azimuthal coordinate. Hence r_{ij} is invariant under the common action of $\mathcal{G}_{\text{yaw}}$ on the token pair. We next construct attention mechanisms that preserve $\mathbf{Z} \mapsto P_k \mathbf{Z}$.

A shared attention family. CERPA restricts positional interactions to the heading-invariant pairwise descriptor in Equation (3). For one attention head of dimension d_h, let $\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i \in \mathbb{R}^{d_h}$ denote the query, key, and value vectors obtained through shared tokenwise projections. We define

$$\ell_{ij} = \frac{\mathbf{q}_i^\top M(r_{ij}) \mathbf{k}_j}{\sqrt{d_h}} + b(r_{ij}), \qquad \mathrm{Attn}(\mathbf{Z})_i = \sum_j \frac{\exp(\ell_{ij})}{\sum_{j'} \exp(\ell_{ij'})} \mathbf{v}_j. \tag{4}$$

Here $\ell_{ij} \in \mathbb{R}$ is the unnormalized attention logit between query token i and key token j, and $\mathrm{Attn}(\mathbf{Z})_i \in \mathbb{R}^{d_h}$ is the output feature at token i for this attention head. Collecting all logits and value vectors, we write $L(\mathbf{Z}) = [\ell_{ij}]$ and $V(\mathbf{Z}) = [\mathbf{v}_1, \dots, \mathbf{v}_{H_t N}]^\top$. The matrix $M(r) \in \mathbb{R}^{d_h \times d_h}$ modulates the query–key interaction, while $b(r) \in \mathbb{R}$ is an additive positional bias (Section 3.3). Both positional terms depend on token positions only through the heading-invariant descriptor r_{ij}.

Proposition 1 (Cyclic-equivariant attention)**.** *Under shared tokenwise projections and the positional interactions in Equation* (4)*, the attention operator is C_N-equivariant:*

$$\mathrm{Attn}(P_k \mathbf{Z}) = P_k\, \mathrm{Attn}(\mathbf{Z}), \qquad g_k \in \mathcal{G}_{\text{yaw}}. \tag{5}$$

Proof. Because r_{ij} is invariant under a common heading shift, the logit matrix satisfies $L(P_k \mathbf{Z}) = P_k L(\mathbf{Z}) P_k^\top$. Shared value projections similarly give $V(P_k \mathbf{Z}) = P_k V(\mathbf{Z})$. Since row-wise softmax commutes with simultaneous row and column permutations, $\mathrm{softmax}(P_k L(\mathbf{Z}) P_k^\top) = P_k\, \mathrm{softmax}(L(\mathbf{Z})) P_k^\top$. Therefore, $\mathrm{Attn}(P_k \mathbf{Z}) = P_k\, \mathrm{Attn}(\mathbf{Z})$. □

Proposition 1 applies to both realizations introduced next: Additive CERPA uses $M(r) = I$ and a circular relative-position bias $b(r)$, whereas Rotary CERPA uses $M(r_{ij}) = R_i^\top R_j$ and $b(r) = 0$.

3.3 Additive and Rotary Realizations

Additive CERPA (CylindricalViT). Additive CERPA instantiates Equation (4) with $M(r) = I, b(r_{ij}) = B_\theta(e_i - e_j, \Delta u_{ij})$, where $\Delta u_{ij} = [u_j - u_i]_N$ and B_θ is a learned circular relative-position bias. The resulting attention logit is

$$\ell_{ij}^{\text{add}} = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_h}} + B_\theta(e_i - e_j, \Delta u_{ij}). \tag{6}$$

The additive form depends only on relative elevation and cyclic azimuth offsets, both invariant under the action of $\mathcal{G}_{\text{yaw}}$. Consequently, the positional bias is preserved under yaw rotation, yielding C_N-equivariant attention.

Rotary CERPA (CylDINO). Rotary CERPA instantiates Equation (4) with $b(r) = 0$ and $M(r_{ij}) = \mathbf{R}_i^\top \mathbf{R}_j$, where $\mathbf{R}_i = \mathbf{R}(e_i, u_i)$ applies a position-dependent rotation to the query and key vectors. As in RoPE (Su et al., 2024), relative position is encoded through the interaction $\mathbf{R}_i^\top \mathbf{R}_j$. Unlike standard RoPE, however, the azimuthal encoding must respect the cyclic domain $\mathbb{Z}_N$. We represent each azimuth column u by its center angle $\phi_u = 2\pi(u + \frac{1}{2})/N$. To ensure exact periodic closure on the cyclic domain, we use integer azimuthal harmonics $m_s \in \mathbb{Z}$. Since $\phi_{u+N} = \phi_u + 2\pi$, integer m_s ensures $\text{Rot}(m_s(\phi_u + 2\pi)) = \text{Rot}(m_s\phi_u)$, so the rotary encoding is identical after one full azimuthal cycle. Using this periodicity, under a heading action $g_k \in \mathcal{G}_{\text{yaw}}$ the azimuth index changes as $u \mapsto [u + k]_N$, giving

$$\begin{aligned} \mathbf{R}(e, [u+k]_N) &= \mathbf{R}_{\text{el}}(e) \oplus \bigoplus_{s=1}^{J} \text{Rot}\left(m_s\phi_u + \frac{2\pi m_s k}{N}\right) \\ &= \left[I_{\text{el}} \oplus \bigoplus_{s=1}^{J} \text{Rot}\left(\frac{2\pi m_s k}{N}\right)\right] \mathbf{R}(e, u). \end{aligned} \tag{7}$$

Here I_{el} denotes the identity on the elevation rotary subspace. The bracketed factor is orthogonal and shared by all tokens for a fixed g_k. Therefore, for any token pair,

$$\mathbf{R}(e_i, [u_i + k]_N)^\top \mathbf{R}(e_j, [u_j + k]_N) = \mathbf{R}_i^\top \mathbf{R}_j. \tag{8}$$

Thus $M(r_{ij}) = \mathbf{R}_i^\top \mathbf{R}_j$ is invariant under the C_N action and depends on azimuth only through $\Delta u_{ij} = [u_j - u_i]_N$. The resulting attention logit is

$$\ell_{ij}^{\text{rot}} = \frac{(\mathbf{R}_i \mathbf{q}_i)^\top (\mathbf{R}_j \mathbf{k}_j)}{\sqrt{d_h}} = \frac{\mathbf{q}_i^\top M(r_{ij}) \mathbf{k}_j}{\sqrt{d_h}}. \tag{9}$$

Rotary CERPA is therefore C_N-equivariant by Proposition 1. We instantiate this construction as CylDINO, a DINOv3-based transformer (Siméoni et al., 2025). To preserve this symmetry, CylDINO uses patch-only attention, excluding class and register tokens. An invariant readout then aggregates the resulting patch features.

3.4 Heading-Invariant RIR Synthesis

A yaw rotation of the panoramic geometry therefore requires both a cyclic shift of the panorama and the corresponding rotation of the 3D XYZ vectors encoded at each pixel. We apply this joint transformation to both the target and reference geometry. For any $g_k \in \mathcal{G}_{\text{yaw}}$, the joint transformation of the geometry induces the same cyclic shift in the CERPA token representation. We then apply an *invariant readout* to these spatial token features, producing a heading-invariant geometry representation. Together with heading-invariant pose conditioning for the target and reference branches, this yields heading-invariant conditioning for the acoustic model.

4 Experiments

Our evaluation addresses four questions: (i) **Heading symmetry:** do the additive and rotary realizations produce heading-equivariant geometric features and heading-invariant RIR predictions? (ii) **Acoustic quality:** does enforcing this symmetry preserve few-shot RIR synthesis quality relative to the original and yaw-augmented models? (iii) **Training efficiency:** does CERPA improve

the performance–model-size trade-off and reduce the training data required to achieve comparable acoustic quality? (iv) **Real-world evaluation:** how well does CERPA synthesize RIRs in real-world acoustic environments? We evaluate these questions on both xRIR and FLAC using their corresponding CERPA realizations.

4.1 Experimental Setup

Dataset and protocol. We evaluate both realizations on **AcousticRooms (AR)** (Liu et al., 2025), following the split of 243 training rooms and 17 held-out rooms used by xRIR and FLAC. We consider both sparse ($K_{\text{ctx}} = 1$) and dense ($K_{\text{ctx}} = 8$) acoustic context. We additionally evaluate sim-to-real transfer by fine-tuning on **Hearing-Anything-Anywhere (HAA)** (Wang et al., 2024), which contains measured RIRs from four environments: a classroom, a dampened room, a hallway, and a complex room. Panoramic depth maps are rendered from scene meshes following FLAC (Brunetto, 2026). Further dataset details are provided in Appendix B.3.

Acoustic metrics. We evaluate RIR synthesis using relative reverberation-time (T60) error, clarity (C50) error, and Early Decay Time (EDT) error for additive CERPA. We additionally report audio-to-audio retrieval recall at ranks 1, 5, and 10 (R@1/5/10) in the geometry-aware AGREE embedding space, measuring whether a synthesized RIR retrieves its corresponding ground-truth RIR for rotary CERPA following FLAC (Brunetto, 2026). Lower T60, C50, and EDT errors indicate better acoustic accuracy, while higher R@1/5/10 indicates better scene consistency.

Training efficiency. We evaluate two forms of efficiency. **Parameter efficiency** measures whether enforcing cyclic equivariance by design improves the performance–model-size trade-off, while **training-data efficiency** measures whether the same inductive bias reduces the training data required to achieve a given performance level. For **parameter efficiency**, we compare CERPA encoders at multiple scales against the largest non-equivariant baseline, keeping the training-data budget fixed. In FLAC, we evaluate CylDINO-S/M/L with 21.6M, 85.7M, and 303.1M encoder parameters against DINOv3-L with 303.1M parameters. In xRIR, we evaluate CylindricalViT-S/M/L with 2.77M, 19.7M, and 43.7M encoder parameters against SimpleViT-L with 43.7M parameters. This setting evaluates whether CERPA can achieve competitive acoustic quality with substantially reduced encoder capacity. For **training-data efficiency**, we fix the encoder scale and use 25%, 50%, 75%, and 100% of the 243 AR training rooms. We construct nested training subsets ($\mathcal{D}_{25} \subset \mathcal{D}_{50} \subset \mathcal{D}_{75} \subset \mathcal{D}_{100}$) to ensure that each larger subset contains all rooms from the smaller subsets. Both studies evaluate on the same complete AR held-out split of 6,337 queries from 17 rooms under both sparse ($K_{\text{ctx}} = 1$) and dense ($K_{\text{ctx}} = 8$) acoustic context using the same five evaluation seeds.

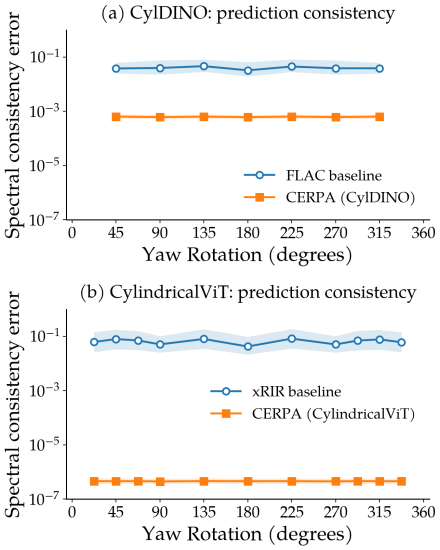

Figure 3: Both CERPA realizations produce nearly unchanged RIR predictions under patch-aligned joint yaw rotations, whereas baseline predictions vary with heading. Lines show median errors, and shaded bands span the 25th to 75th percentiles.

Baselines. We evaluate two realizations of CERPA to cover deterministic and probabilistic RIR synthesis. For deterministic prediction, we integrate Additive CERPA, implemented as CylindricalViT, into xRIR (Liu et al., 2025). For probabilistic generation, we integrate Rotary CERPA, implemented as CylDINO, into FLAC (Brunetto, 2026). For xRIR, we compare the original xRIR, Yaw-Augmented xRIR, and Additive CERPA (CylindricalViT). For FLAC, we compare the original FLAC, Yaw-Augmented FLAC, and Rotary CERPA (CylDINO). The yaw-augmented baselines apply joint yaw augmentation to the panoramic geometry and associated source–receiver coordinates during training, but do not impose heading equivariance architecturally.

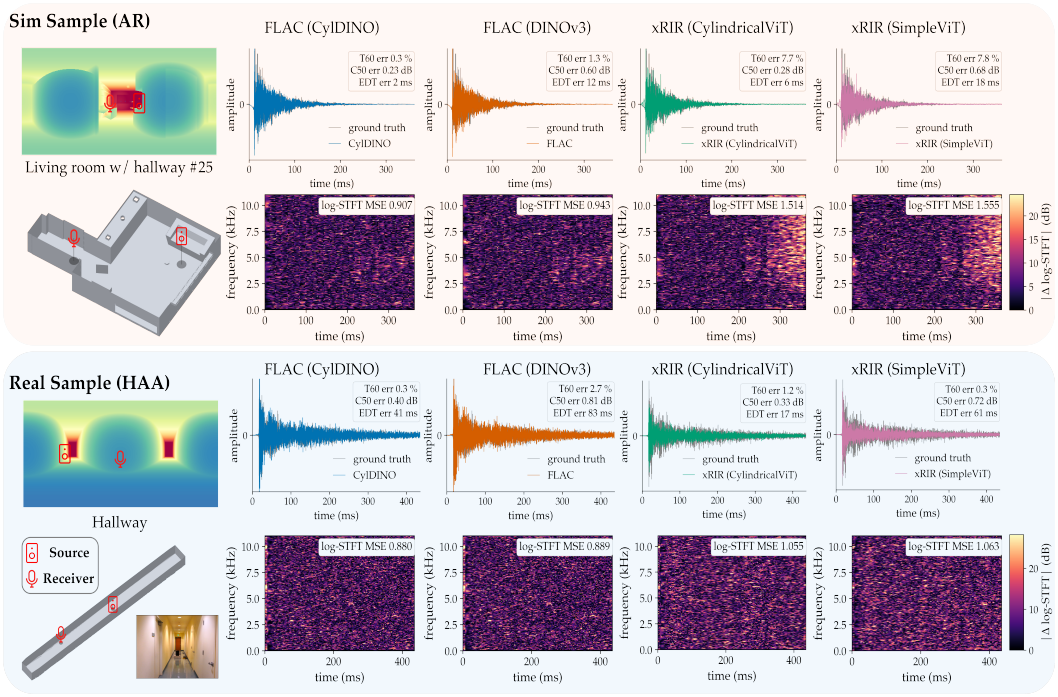

Figure 4: **Qualitative comparison on a simulated and a real room.** Top: AR unseen room *Living room w/ hallway 25*. Bottom: HAA *hallway*. Every model receives $K{=}8$ reference measurements. The panorama depth images are rendered at the receiver (AR) or at the loudspeaker (HAA); First row per sample: predicted RIR (colour) over the ground truth (grey); Second row: $|\Delta \log|\mathrm{STFT}||$ between prediction and ground truth in dB.

4.2 Results and Discussions

Heading equivariance and invariance. Figure 3 evaluates RIR prediction consistency under patch-aligned joint yaw rotations of panoramic geometry and source–receiver poses on the AR unseen split. Compared with their respective baselines, CylDINO and CylindricalViT reduce median yaw-induced mean absolute log-spectral differences by nearly two and approximately five orders of magnitude, respectively. These results demonstrate CERPA's nearly yaw-invariant RIR predictions. Appendix C.2 provides further evaluation details.

Acoustic synthesis quality. Enforcing heading symmetry yields acoustic synthesis quality better than the baselines. For xRIR, we report T60, C50, and EDT errors in seen and unseen environments. For FLAC, we additionally report AGREE retrieval recall at ranks 1, 5, and 10. Figure 4 presents qualitative examples of both realizations and their respective baselines on AR and HAA.

Additive CERPA *in xRIR.* The upper block of Table 1 compares xRIR, Yaw-Augmented xRIR, and CylindricalViT under $K_{\mathrm{ctx}} \in \{1, 8\}$. With dense context ($K_{\mathrm{ctx}} = 8$), CylindricalViT achieves the lowest C50 and EDT errors on both seen and unseen splits, outperforming both baselines. With sparse context ($K_{\mathrm{ctx}} = 1$), CylindricalViT incurs modest increases in all three errors relative to the original xRIR on both splits. Overall, Additive CERPA maintains competitive acoustic quality while enforcing encoder equivariance, with its clearest benefits in clarity and early-decay accuracy under dense context.

Rotary CERPA *in FLAC.* Under the same training budget, CylDINO improves five of the six reported metrics over the original FLAC under both sparse and dense acoustic context. Across all three variants, it achieves the lowest EDT error and the highest R@1 and R@10 at both $K_{\mathrm{ctx}} = 1$ and 8. It also achieves the highest R@5 at $K_{\mathrm{ctx}} = 1$. Although these gains come with higher C50 error, they demonstrate that Rotary CERPA improves early-decay accuracy and acoustic retrieval while enforcing C_N-equivariance by construction.

Table 1: **RIR synthesis performance under sparse ($K_{\mathrm{ctx}} = 1$) and dense ($K_{\mathrm{ctx}} = 8$) acoustic context.** Results are reported as mean ± standard deviation across five evaluation seeds (42–46).

Additive CERPA

Method	K_{ctx}	Seen split			Unseen split		
		T60 (%) ↓	C50 (dB) ↓	EDT (ms) ↓	T60 (%) ↓	C50 (dB) ↓	EDT (ms) ↓
xRIR	1	$10.61_{\pm 0.21}$	$1.490_{\pm 0.017}$	$57.5_{\pm 0.6}$	$\mathbf{13.38}_{\pm 0.10}$	$\mathbf{1.890}_{\pm 0.013}$	$\mathbf{68.3}_{\pm 0.6}$
YAW-AUGMENTED XRIR	1	$\mathbf{10.27}_{\pm 0.21}$	$\mathbf{1.472}_{\pm 0.014}$	$\mathbf{56.7}_{\pm 0.7}$	$13.39_{\pm 0.08}$	$1.969_{\pm 0.015}$	$69.5_{\pm 0.7}$
XRIR (CYLINDRICALVIT)	1	$11.12_{\pm 0.26}$	$1.527_{\pm 0.012}$	$60.6_{\pm 0.6}$	$13.58_{\pm 0.09}$	$1.942_{\pm 0.014}$	$69.8_{\pm 0.8}$
xRIR	8	$6.89_{\pm 0.02}$	$0.908_{\pm 0.005}$	$34.4_{\pm 0.2}$	$9.58_{\pm 0.01}$	$1.238_{\pm 0.005}$	$47.0_{\pm 0.2}$
YAW-AUGMENTED XRIR	8	$\mathbf{6.84}_{\pm 0.01}$	$0.917_{\pm 0.005}$	$35.1_{\pm 0.1}$	$\mathbf{9.32}_{\pm 0.01}$	$1.294_{\pm 0.006}$	$46.4_{\pm 0.1}$
XRIR (CYLINDRICALVIT)	8	$6.92_{\pm 0.02}$	$\mathbf{0.890}_{\pm 0.005}$	$\mathbf{34.1}_{\pm 0.3}$	$9.44_{\pm 0.01}$	$\mathbf{1.231}_{\pm 0.006}$	$\mathbf{45.1}_{\pm 0.3}$

Rotary CERPA

Method	K_{ctx}	EDT (ms) ↓	C50 (dB) ↓	T60 (%) ↓	R@1 (%) ↑	R@5 (%) ↑	R@10 (%) ↑
FLAC	1	$43.437_{\pm 0.397}$	$1.0884_{\pm 0.0088}$	$10.287_{\pm 0.026}$	$4.990_{\pm 0.115}$	$15.111_{\pm 0.129}$	$22.509_{\pm 0.181}$
Yaw-Augmented FLAC	1	$42.447_{\pm 0.379}$	$\mathbf{1.0855}_{\pm 0.0051}$	$\mathbf{9.397}_{\pm 0.052}$	$5.031_{\pm 0.063}$	$15.919_{\pm 0.256}$	$23.166_{\pm 0.135}$
FLAC (CYLDINO)	1	$\mathbf{41.788}_{\pm 0.377}$	$1.1439_{\pm 0.0021}$	$9.847_{\pm 0.052}$	$\mathbf{5.419}_{\pm 0.156}$	$\mathbf{15.979}_{\pm 0.127}$	$\mathbf{23.604}_{\pm 0.238}$
FLAC	8	$40.650_{\pm 0.101}$	$\mathbf{1.0093}_{\pm 0.0035}$	$8.993_{\pm 0.011}$	$5.173_{\pm 0.138}$	$15.430_{\pm 0.197}$	$23.409_{\pm 0.056}$
Yaw-Augmented FLAC	8	$39.923_{\pm 0.057}$	$1.0132_{\pm 0.0016}$	$\mathbf{7.965}_{\pm 0.014}$	$5.391_{\pm 0.097}$	$\mathbf{16.563}_{\pm 0.169}$	$23.961_{\pm 0.107}$
FLAC (CYLDINO)	8	$\mathbf{39.076}_{\pm 0.117}$	$1.0786_{\pm 0.0021}$	$8.313_{\pm 0.021}$	$\mathbf{5.479}_{\pm 0.111}$	$16.342_{\pm 0.137}$	$\mathbf{24.226}_{\pm 0.121}$

Training efficiency. Tables 2 and 3 examine parameter and training-data efficiency, respectively.

Parameter efficiency. At $K_{\mathrm{ctx}} = 8$, CYLINDRICALVIT-S achieves lower EDT and C50 errors than xRIR-L, with comparable T60 error, using only 2.78M versus 43.71M encoder parameters (93.6% reduction). CYLINDRICALVIT-M outperforms xRIR-L on all three acoustic metrics and achieves the lowest EDT and C50 errors among the evaluated additive variants. For the rotary realization, CYLDINO-M improves R@1, R@5, and R@10 over DINOv3-L by approximately 9.1%, 5.3%, and 5.7%, respectively, while using 85.7M versus 303.1M encoder parameters (71.7% reduction). CYLDINO-L achieves the highest R@5 and R@10 among the compared encoders. Together, these results demonstrate parameter-efficiency gains in acoustic reconstruction for the additive realization and retrieval for the rotary realization.

Training-data efficiency. CYLINDRICALVIT improves C50 and T60 errors over SimpleViT at every evaluated data fraction, and improves all three metrics at 50%, 75%, and 100%. For the rotary realization, CYLDINO outperforms DINOv3 on all three metrics using 25% of the training data, reducing EDT and T60 errors by approximately 3.0% and 12.1%, respectively. At 50% and 75%, it reduces C50 error by approximately 0.4% and 3.2%, respectively. With the full dataset, it improves EDT and T60. These results demonstrate benefits under limited training data. Additional results are shown in Appendix C.

Sim-to-real transfer. We fine-tune xRIR, FLAC, their yaw-augmented baselines, and their CERPA counterparts on HAA, using orientation-aware conditioning for the CERPA variants. Table 4 summarizes their performance on measured real-world RIRs. For the additive realization, CYLINDRICALVIT achieves the lowest EDT and C50 errors in the classroom, the lowest errors on all three metrics in the hallway, and the lowest EDT and T60 errors in the complex room. These improvements hold against both the original and yaw-augmented baselines. For the rotary realization, CYLDINO reduces C50

Table 2: **Parameter efficiency of CERPA (AcousticRooms unseen split, $K_{\mathrm{ctx}} = 8$).** We compare (a) CYLINDRICALVIT-S/M/L against xRIR-L (12 epochs) and (b) CYLDINO-S/M/L against DINOv3-L (15k updates). Results are means over five evaluation seeds (42–46). T60, C50, and EDT are measured in %, dB, and ms, respectively; recall is reported in %.

(a) Additive CERPA (acoustic model backbone: xRIR)

Encoder	T60 ↓	C50 ↓	EDT ↓
SimpleViT-L	9.54	1.275	47.13
CYLINDRICALVIT-S	9.64	1.260	46.38
CYLINDRICALVIT-M	9.44	**1.231**	**45.06**
CYLINDRICALVIT-L	**9.37**	1.243	46.08

(b) Rotary CERPA (acoustic model backbone: FLAC)

Encoder	R@1 ↑	R@5 ↑	R@10 ↑
DINOv3-L	3.62	12.57	19.41
CYLDINO-S	2.49	8.98	14.71
CYLDINO-M	**3.95**	13.23	20.51
CYLDINO-L	3.90	**13.56**	**20.96**

Table 3: **Training-data efficiency of CERPA on AcousticRooms.** We vary the training-data fraction, using (a) 12 epochs for xRIR and (b) 40k updates for FLAC at $K_{\text{ctx}} = 8$. The xRIR panel reports means; FLAC reports mean $\pm$ standard deviation over five evaluation seeds (42–46). EDT, C50, and T60 are measured in ms, dB, and percent.

(a) Additive CERPA (acoustic model backbone: xRIR)

Data	Encoder	EDT ↓	C50 ↓	T60 ↓
25%	SimpleViT	**48.250**	1.362	9.425
25%	CylindricalViT	48.860	**1.352**	**8.939**
50%	Simple	48.050	1.336	9.661
50%	CylindricalViT	**47.250**	**1.307**	**9.489**
75%	SimpleViT	47.240	1.263	9.575
75%	CylindricalViT	**47.020**	**1.261**	**9.391**
100%	SimpleViT	47.020	1.238	9.575
100%	CylindricalViT	**45.060**	**1.231**	**9.443**

(b) Rotary CERPA ($K_{\text{ctx}} = 8$) (acoustic model backbone: FLAC)

Data	Encoder	EDT ↓	C50 ↓	T60 ↓
25%	DINOv3	$42.914_{\pm 0.058}$	$1.048_{\pm 0.002}$	$10.253_{\pm 0.012}$
25%	CylDINO	$\mathbf{41.613}_{\pm 0.081}$	$\mathbf{1.044}_{\pm 0.001}$	$\mathbf{9.013}_{\pm 0.017}$
50%	DINOv3	$\mathbf{41.588}_{\pm 0.064}$	$1.044_{\pm 0.002}$	$\mathbf{9.450}_{\pm 0.010}$
50%	CylDINO	$42.792_{\pm 0.056}$	$\mathbf{1.040}_{\pm 0.003}$	$9.500_{\pm 0.015}$
75%	DINOv3	$\mathbf{41.640}_{\pm 0.060}$	$1.104_{\pm 0.002}$	$\mathbf{8.913}_{\pm 0.009}$
75%	CylDINO	$42.270_{\pm 0.050}$	$\mathbf{1.069}_{\pm 0.002}$	$9.054_{\pm 0.012}$
100%	DINOv3	$40.650_{\pm 0.101}$	$\mathbf{1.009}_{\pm 0.004}$	$8.993_{\pm 0.011}$
100%	CylDINO	$\mathbf{39.076}_{\pm 0.117}$	$1.079_{\pm 0.002}$	$\mathbf{8.313}_{\pm 0.021}$

Table 4: **Sim-to-real RIR synthesis on Hearing-Anything-Anywhere (HAA).** Additive CERPA reports per-environment acoustic errors, whereas Rotary CERPA adds retrieval recall. Units are EDT (s for Additive, ms for Rotary), C50 (dB), and T60 and recall (%).

(a) Additive CERPA ($K_{\text{ctx}} = 8$)

Method	Classroom			Dampened Room			Hallway			Complex Room		
	EDT	C50	T60	EDT	C50	T60	EDT	C50	T60	EDT	C50	T60
xRIR	0.096	1.97	**4.32**	0.038	**2.75**	-	0.095	1.70	4.42	0.107	2.23	4.96
Yaw-Augmented xRIR	0.081	1.52	4.33	**0.037**	2.79	-	0.071	1.41	3.14	0.081	**1.74**	5.05
xRIR (CylindricalViT)	**0.079**	**1.50**	4.52	0.038	2.88	-	**0.062**	**1.14**	**2.72**	**0.080**	1.75	**4.84**

(b) Rotary CERPA

Method	K_{ctx}	T60	C50	EDT	R@1	R@5	R@10
FLAC	1	**3.617**	2.254	**87.95**	**5.08**	19.07	31.22
Yaw-Augmented FLAC	1	4.238	2.775	93.89	4.04	16.27	27.20
FLAC (CylDINO)	1	3.680	**2.231**	90.99	4.68	**19.16**	**31.44**
FLAC	8	**3.413**	2.202	**84.99**	**5.18**	**19.17**	**31.69**
Yaw-Augmented FLAC	8	4.092	2.777	91.76	4.13	16.10	27.34
FLAC (CylDINO)	8	3.553	**2.158**	85.82	4.96	18.77	31.45

error relative to FLAC by approximately 1.0% ($K_{\text{ctx}} = 1$) and 2.0% ($K_{\text{ctx}} = 8$) and achieves the highest R@5 and R@10 ($K_{\text{ctx}} = 1$). It also outperforms Yaw-Augmented FLAC on all six metrics at both context sizes on HAA fine-tuning. Overall, the adapted CERPA variants achieve better real-world synthesis quality. The fine-tuning protocol and orientation-aware adaptations are detailed in Appendix C.3.

5 Conclusion

We introduced CERPA, a unified framework for heading-equivariant panoramic attention in few-shot room-acoustic synthesis. Its additive and rotary realizations encode cyclic positional structure through circular relative-position biases and integer-harmonic rotary embeddings, yielding CylindricalViT for xRIR and CylDINO for FLAC. Together with consistent geometric transformations, invariant readout, and heading-invariant pose conditioning, they enforce heading-invariant acoustic conditioning under patch-aligned yaw transformations. Experiments on AcousticRooms demonstrate better heading consistency and improvements over the original and yaw-augmented baselines on key acoustic and retrieval metrics. Sim-to-real transfer results on HAA show improved acoustic accuracy across multiple environments with CylindricalViT and CylDINO relative to their respective baselines. Future work may explore hybrid CERPA realizations that jointly learn rotary interactions $M(r_{ij})$ and biases $b(r_{ij})$. Additional directions include extending the framework beyond patch-aligned yaw rotations and improving its transfer to diverse real-world acoustic environments.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, design or provide feedback on research methodology or experiments, clean and reformat dataset and implement methods. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses and interpret results. And generate synthetic data sets, assist with translation and support qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code, creation of artifacts, draft parts of a research paper, summarize or analyse existing literature, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature and format references. We have reviewed all AI-assisted work. The authors reviewed the AI-assisted material, checked numerical statements against experimental results, verified references against the cited publications, and inspected generated figures and scripts for consistency with the data and evaluation protocols. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe the experimental settings and implementation details to support reproducibility, and provide the full implementation of our method in an anonymized repository linked in Appendix A. The problem setup, heading symmetry requirements, and unified attention framework are stated in Section 3, with the formal proof for cyclic-equivariant attention provided in Section 3.2. Section 3.3 and Section 3.4 describe the additive and rotary realizations alongside the invariant acoustic conditioning, while Appendix B.1 details the network architectures, training objectives, and inference procedures. Architecture and optimization hyperparameters are thoroughly summarized in Table 5 for CYLDINO and Table 6 for CYLINDRICALVIT. Section 4.1 describes the dataset splits, acoustic context sizes, baselines, and evaluation protocol. Appendix B.2 specifies the baseline implementations, including exact yaw-augmentation procedures, while Appendix B.3 documents the benchmark dataset processing. Finally, Appendices C.3, C.4, and C.5 detail the fine-tuning protocols, optimization parameters, and orientation-aware cylindrical conditioning required to reproduce the real-world acoustic transfer experiments on HAA.

REFERENCES

Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, Jiankang Deng, and Xiatian Zhu. AV-GS: Learning material and geometry aware priors for novel view acoustic synthesis. In *Advances in Neural Information Processing Systems*, 2024.

Amandine Brunetto. Few-shot acoustic synthesis with multimodal flow matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15773–15783, 2026.

Amandine Brunetto, Sascha Hornauer, and Fabien Moutarde. NeRAF: 3D scene infused neural radiance and acoustic fields. In *International Conference on Learning Representations*, 2025.

Mingfei Chen and Eli Shlizerman. AV-Cloud: Spatial audio rendering through audio-visual cloud splatting. In *Advances in Neural Information Processing Systems*, 2024.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, 2016.

Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018.

Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. SphereNet: Learning spherical representations for detection and classification in omnidirectional images. In *European Conference on Computer Vision*, pp. 525–541, 2018.

Peijian Ding, Davit Soselia, Thomas Armstrong, Jiahao Su, and Furong Huang. Reviving shift equivariance in vision transformers. In *ICML Workshop on Spurious Correlations, Invariance and Stability*, 2023.

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning SO(3) equivariant representations with spherical CNNs. In *European Conference on Computer Vision*, 2018.

Haoran Feng, Dizhe Zhang, Xiangtai Li, Bo Du, and Lu Qi. DiT360: High-fidelity panoramic image generation via hybrid training. *arXiv preprint arXiv:2510.11712*, 2025.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, 2024.

Or Hirschorn, Aaron Olender, Eli Alshan, Ianir Ideses, Lior Fritz, and Sagie Benaim. SpheRoPE: Zero-shot optimization-free 360 panorama generation with spherical RoPE. *arXiv preprint*, 2026.

Diwei Huang, Kunyang Lin, Peihao Chen, and Qing Du. Map-guided few-shot audio-visual acoustics modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025.

Michael J. Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. LieTransformer: Equivariant self-attention for lie groups. In *International Conference on Machine Learning*, 2021.

Jungwoon Kang, Jaehun Kim, Yiwon Yu, Hyungyum Jang, Sanghoon Lee, and Jongyoo Kim. SHERPA: Seam-aware harmonized ERP adaptation for open-domain 360° panorama generation. *arXiv preprint*, 2026.

Konstantinos Kogkalidis, Jean-Philippe Bernardy, and Vikas Garg. Algebraic positional encodings. In *Advances in Neural Information Processing Systems*, 2024.

Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. In *Advances in Neural Information Processing Systems*, 2024.

Tobias Lentz, Dirk Schröder, Michael Vorländer, and Ingo Assenmacher. Virtual reality system with integrated sound field simulation and reproduction. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. AV-NeRF: Learning neural fields for real-world audio-visual scene synthesis. In *Advances in Neural Information Processing Systems*, 2023a.

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

Zhixin Ling, Zhen Xing, Xiangdong Zhou, Manliang Cao, and Guichun Zhou. PanoSwin: A pano-style swin transformer for panorama understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Andrew F. Luo, Yilun Du, Michael J. Tarr, Joshua B. Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3165–3177, 2022.

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

Sophie Ostmeier, Brian Axelrod, Maya Varma, Michael Moseley, Akshay S. Chaudhari, and Curtis Langlotz. LieRE: Lie rotational positional encodings. In *International Conference on Machine Learning*, 2025.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Renan A. Rojas-Gomez, Teck-Yian Lim, Minh N. Do, and Raymond A. Yeh. Making vision transformers truly shift-equivariant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5568–5577, 2024.

David W. Romero and Jean-Baptiste Cordonnier. Group equivariant stand-alone self-attention for vision. In *International Conference on Learning Representations*, 2021.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. PanoFormer: Panorama transformer for indoor 360° depth estimation. In *European Conference on Computer Vision*, pp. 195–211, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, et al. DINOv3. *arXiv preprint*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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

David Thery and Brian FG Katz. Anechoic audio and 3D-video content database of small ensemble performances for virtual concerts. In *International Congress on Acoustics*, 2019.

Michael Vorländer. *Auralization*. Springer, 2020.

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

Maurice Weiler and Gabriele Cesa. General E(2)-equivariant steerable CNNs. In *Advances in Neural Information Processing Systems*, 2019.

Zihan Weng, Chanlin Yi, Pouya Bashivan, Jing Lu, Fali Li, Dezhong Yao, Jingming Hou, Yangsong Zhang, et al. SPECTRE: Spectral pre-training embeddings with cylindrical temporal rotary position encoding for fine-grained sEMG-based movement decoding. *arXiv preprint*, 2025.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Yifan Zhang, Zixiang Chen, Yifeng Liu, Zhen Qin, Huizhuo Yuan, Kangping Xu, Yang Yuan, Quanquan Gu, et al. Group representational position encoding. In *International Conference on Learning Representations*, 2026.

A CODE AVAILABILITY

To facilitate the review process, we provide the full implementation of our method in an anonymized repository: `https://anonymous.4open.science/r/CERPA-review/`. Instructions for setting up and reproducing the proposed method are provided in the repository `README`.

B EXPERIMENTAL DETAILS

We train the CYLDINO-family models for 40,000 optimization steps using distributed data parallelism on two NVIDIA A6000 GPUs. The per-GPU batch size is 32, giving a global batch size of 64 without gradient accumulation. Checkpoints are evaluated every 2,500 steps, while the reported matched-budget comparisons use the final checkpoint at 40,000 steps. We train the CYLINDRICALVIT-family models for 12 epochs with an effective batch size of 64 using a single NVIDIA RTX A6000 GPU (48 GB) and evaluate their final checkpoints (about $12 \times 4,631 = 55,572$ steps) on AR dataset. Using this training and evaluation protocol, we compare the models under matched training budgets in the following experiments.

B.1 CERPA ARCHITECTURE AND TRAINING CONFIGURATION

We evaluate the additive and rotary realizations of CERPA within xRIR and FLAC, respectively. Within each realization, we keep the acoustic-generation architecture and optimization settings fixed across variants to isolate the effect of geometry and pose conditioning. The following sections describe the configurations used for the main acoustic-synthesis comparisons in Table 1.

CYLDINO architecture and training configuration. We initialize the geometry encoders of FLAC and CYLDINO from the same official DINOv3 ViT-S/16 checkpoint and fine-tune them for downstream RIR synthesis without additional domain-specific self-supervised pretraining. CYLDINO combines cylindrical coordinate preprocessing, integer-harmonic azimuthal RoPE, mean pooling over patch tokens, and heading-invariant pose encoding. The CLS and register tokens are removed, and the geometry backbone is shared across source and context-pose conditioning. In contrast, FLAC retains its original geometry and pose conditioning. These modifications enforce heading consistency under patch-aligned yaw transformations without averaging over rotated scene representations.

Controlled optimization. The FLAC-family models use the same training seed (42), data order, optimizer, learning-rate schedule, batch size, and optimization budget, but are optimized independently. We train each model for 40,000 optimization steps using distributed data parallelism on two NVIDIA A6000 GPUs. A per-GPU batch size of 32 gives a global batch size of 64 without gradient accumulation. Checkpoints are evaluated every 2,500 steps, and all main comparisons use the final checkpoint at 40,000 steps.

Distributed training and memory management. Synchronized Batch Normalization aggregates statistics across both workers, yielding a normalization batch size of 64. We retain this configuration across variants because gradient accumulation alone does not reproduce the same BatchNorm statistics. Activation checkpointing is applied to the geometry backbone using PyTorch's non-reentrant implementation (`use_reentrant=False`), with the same setting used for the baseline. Training uses bfloat16 mixed precision, and the pretrained VAE remains frozen.

Evaluation protocol. We evaluate sparse ($K_{\mathrm{ctx}} = 1$) and dense ($K_{\mathrm{ctx}} = 8$) acoustic context using one sampling step, a classifier-free guidance scale of 1.0, and evaluation seeds 42–46. Reported standard deviations reflect evaluation-seed variability, not variation across independently trained models. Table 5 summarizes the architecture, training, and evaluation settings.

CYLINDRICALVIT architecture and training configuration. The original xRIR (Liu et al., 2025) models use a SimpleViT-based geometry encoder with 12 transformer layers, a hidden dimension of 512, and eight attention heads. We follow the released implementation's 12-layer configuration rather than the six-layer architecture described in the original paper. The encoder processes a three-channel XYZ panorama of resolution 256×512 using 16×32 patches, producing a 16×16 token grid. CYLINDRICALVIT introduces a per-column cylindrical coordinate transformation, elevation-only

Table 5: **Model architecture and training configuration of CylDINO.** Settings correspond to the model evaluated after 40,000 optimization steps in Table 1. The geometry backbone is initialized from the official DINOv3 ViT-S/16 checkpoint and fine-tuned for RIR synthesis without additional domain-specific self-supervised pretraining.

Category	Parameter	Configuration
Architecture	Geometry encoder	Cylindrical DINOv3 ViT-S/16; 21.6M backbone parameters
	Backbone initialization	`dinov3-vits16-pretrain-lvd1689m`
	Backbone sharing	Shared by source and context-pose geometry conditioners
	Generator	FLAC latent rectified-flow DiT
	DiT depth / hidden dimension / heads	12 / 256 / 8
	Cross-attention / global conditioning	256-dimensional context tokens / 512-dimensional AdaLN conditioning
	Latent input/output channels	32
	VAE pretransform	Pretrained and frozen
Input	Geometry representation	Ego-centred XYZ displacement field, $\mathbf{x}_{\text{query}} - \mathbf{x}_{\text{surface}}$
	Input dimensions / patch size	$3 \times 256 \times 512$ / 16×16 pixels
	Patch grid / number of patch tokens	16×32 / 512
	RIR channels / sampling rate	Single-channel / 22,050 Hz
	Waveform length	10,240 samples (≈ 0.464 s)
Conditioning	Cylindrical XYZ gauge	Per-column rotation of horizontal XYZ components into the local cylindrical basis
	Azimuthal positional encoding	Integer-harmonic RoPE, $m \in \{0, \ldots, 15\}$
	Prefix tokens / geometry readout	CLS and register tokens removed / mean over patch tokens
	Pose representation	Heading-invariant encoding of source–receiver geometry
	Additional backbone parameters	None
Training	Hardware	2× NVIDIA A6000 GPUs, DDP
	Optimizer	AdamW; learning rate 5×10^{-5}, $\beta = (0.9, 0.999)$, weight decay 10^{-3}
	Training budget	40,000 optimization steps
	Per-GPU batch / gradient accumulation	32 / none
	Global batch / normalization	64 / synchronized BatchNorm
	Learning-rate schedule	InverseLR; `inv_gamma` $= 10^6$, `power` $= 0.5$, `warmup` $= 0.99$
	Training objective	Rectified-flow objective in the VAE latent space
	Conditioning dropout probability	0.1
	Precision / training seed	BF16 mixed precision / 42
	Backbone activation checkpointing	Enabled
Evaluation	Acoustic context size	$K_{\text{ctx}} \in \{1, 8\}$
	Test-time frame averaging	None
	Conditioning precision	BF16 autocast
	Sampling steps / CFG scale	1 / 1.0
	Evaluation seeds	42–46; results reported as mean ± standard deviation

positional encoding, and circular relative-position biases. The ResNet-18 audio encoder and the remaining acoustic-generation components are unchanged. Detailed descriptions of configurations and hyperparameters of the training and evaluation of CylindricalViT are shown in Table 6.

Controlled optimization. We train the xRIR-family models for 12 epochs on a single NVIDIA A6000 GPU. The per-step batch size is 32, with two gradient-accumulation steps giving an effective batch size of 64. All variants use AdamW with an initial learning rate of 10^{-3} and weight decay of 10^{-4}. The learning rate is reduced by a factor of 10 every three epochs. The training objective combines magnitude-STFT L_1 loss with energy-decay L_1 loss weighted by 0.01. We keep the training data, optimization settings, and training budget fixed across variants and use the final-epoch checkpoints for evaluation.

B.2 Few-Shot Acoustic synthesis Baseline Implementation

Yaw-Augmented FLAC. This baseline retains FLAC's architecture and encourages yaw robustness through training-time data augmentation. At each optimizer step, every training example independently draws an integer offset $d_i \sim \mathrm{Unif}\{0, \ldots, 511\}$, corresponding to $\alpha_i = 2\pi d_i/512$ (0.703125° per column). However, for each training example, we randomly sample one of the 512 panorama-aligned rotations. The equirectangular depth point cloud is circularly rolled by d_i columns, and its per-pixel Cartesian vectors together with all source and context pose fields are rotated by the corresponding yaw angle α_i about the vertical axis. The monaural target RIR and the context RIRs are left unchanged, since this common yaw transformation is target-preserving under the symmetry in Eq. 1. Yaw offsets are generated deterministically from a base seed of 42 for reproducibility. No random yaw augmentation is used at inference.

Yaw-Augmented xRIR. This baseline retains xRIR's original SimpleViT geometry encoder, acoustic encoder, and prediction head, and introduces random yaw augmentation during training. For each training example in every mini-batch, we independently sample an integer offset $d_i \sim \mathrm{Unif}\{0, \ldots, 511\}$, corresponding to a yaw angle $\alpha_i = 2\pi d_i/512$ (0.703125° per column). The equirectangular depth point cloud is circularly rolled by d_i columns, and its per-pixel Cartesian vectors, the target-source position, and all reference-source positions are jointly rotated by α_i about the vertical axis. All positions are expressed relative to the receiver, which remains at the origin. The monaural target and reference RIRs are left unchanged, since this common yaw transformation preserves the acoustic target under the symmetry in Eq. 1. Yaw offsets are generated reproducibly from a base seed of 0 and the mini-batch counter. The model uses the same training objective and 12-epoch budget as the unaugmented xRIR baseline. No random yaw augmentation is used at inference.

B.3 Benchmark & Dataset Selection

We evaluate on AcousticRooms (AR) and Hearing-Anything-Anywhere (HAA) to study cross-room generalization in simulation and transfer to measured real-world acoustics, respectively. Both datasets provide monaural RIRs, source–receiver locations, and geometric information suitable for panoramic conditioning. Together, they allow us to examine heading consistency and acoustic synthesis quality across diverse simulated scenes and real recording conditions.

AcousticRooms (AR). AcousticRooms (Liu et al., 2025) contains approximately 300,000 simulated RIRs spanning 260 rooms across ten categories: apartments, auditoriums, bathrooms, bedrooms, cafes, listening rooms, living rooms, meeting rooms, offices, and restaurants. The dataset is generated using the Treble acoustic simulator, combining wave-based simulation at low frequencies with geometric simulation at high frequencies. Room surfaces receive randomized acoustic properties drawn from 332 materials across 11 material categories, allowing geometrically similar rooms to exhibit different acoustic responses. Sources and receivers are distributed in three dimensions, with omnidirectional sources and monaural receivers.

Each RIR is associated with source–receiver coordinates and a 256×512 equirectangular depth panorama rendered at the receiver position. We project the depth panorama into three-dimensional coordinates to construct the geometry inputs used by both host systems. The simulation meshes are untextured, so geometry conditioning uses depth-derived coordinates rather than RGB appearance. RIRs are sampled at 22,050 Hz; waveform lengths and model-specific preprocessing are specified in Tables 6 and 5.

We use 243 rooms for training and 17 held-out rooms for unseen-room evaluation. Our unseen evaluation split contains 6,337 target queries and is kept fixed across model variants. In the few-shot setting, contextual RIRs share the target receiver but correspond to different source positions. We evaluate both sparse ($K_{\mathrm{ctx}} = 1$) and dense ($K_{\mathrm{ctx}} = 8$) acoustic context. Within each host system, all variants use the same evaluation queries and context-selection protocol.

Hearing-Anything-Anywhere (HAA). HAA (Wang et al., 2024) provides measured monaural RIRs from four real environments: a classroom, a dampened room, a hallway, and a complex room. Each environment contains a fixed sound source and measurements at multiple receiver positions. Unlike the simulated omnidirectional sources in AR, the real recordings include source directivity and measurement noise, introducing an acoustic domain shift.

The recordings are originally sampled at 48 kHz and are resampled to 22,050 Hz for compatibility with the host systems. HAA does not directly provide panoramic depth maps. Following FLAC (Brunetto, 2026), we reconstruct room meshes from the provided wall and surface annotations and render panoramic depth maps at the fixed source positions. These simplified geometric annotations provide less detailed scene structure than the simulated geometry in AR, introducing an additional domain gap in the conditioning inputs.

HAA also differs from AR in its spatial sampling: contextual measurements share a source rather than a receiver. This distinction determines the panorama location and the organization of source–receiver conditioning. We use HAA to assess whether the proposed conditioning remains effective under real acoustic measurements and approximate scene geometry. Within each host system, baseline and CERPA variants are evaluated using the same data partition, geometric inputs, and acoustic-context protocol.

C Additional RIR Synthesis Results

C.1 Pilot Study: Sensitivity to the Heading Reference

We test whether a few-shot RIR synthesis model depends on the choice of global heading reference. A global yaw rotation changes only the coordinate representation: the physical room, source, and receiver remain fixed, so the target monaural RIR should not change.

Experimental setup. We evaluate the released, frozen FLAC checkpoint on the full AcousticRooms unseen split with $K_{\mathrm{ctx}} = 1$ (6,337 queries from 17 rooms). For each query, we jointly rotate the conditioning panorama and all geometric inputs by $\alpha \in 0°, 90°, 180°, 270°$. The equirectangular geometry is rolled by the corresponding azimuth offset, while its per-pixel 3D vectors and all source–receiver pose vectors are rotated about the vertical axis by the same angle. Reference RIRs and the ground-truth target RIR are unchanged. All runs use the same checkpoint, random seed (42), sampling noise, and inference settings. Figure 5 shows the transformation. The radial-distance panorama is cyclically shifted, while jointly rotating the vector values and their image locations preserves the underlying geometry. The resulting inputs therefore describe the same physical scene under different heading references.

Yaw rotation changes the predicted RIR. We measure the invariance gap by comparing the prediction under rotated conditioning, P_α, with the original prediction P_0. As shown in Table 7(b), changing only the heading reference produces a relative waveform ℓ_2 difference of 19–22%. The predictions also differ in their acoustic properties: the mean EDT gap is 18.65–20.11,ms, the T60 gap is 3.3–3.4%, and the C50 gap is 0.55–0.60,dB. All of these gaps would be zero for a heading-invariant predictor.

Figure 6 shows one example at the 90th percentile of the per-sample envelope-gap distribution. The ground-truth RIR is unchanged, but a 180° rotation produces a visibly different predicted waveform and decay envelope. The waveform relative-ℓ_2 difference for this example is 0.278. The difference trace and separated Schroeder decay curves show that the discrepancy is not explained by a small temporal shift or a localized waveform perturbation.

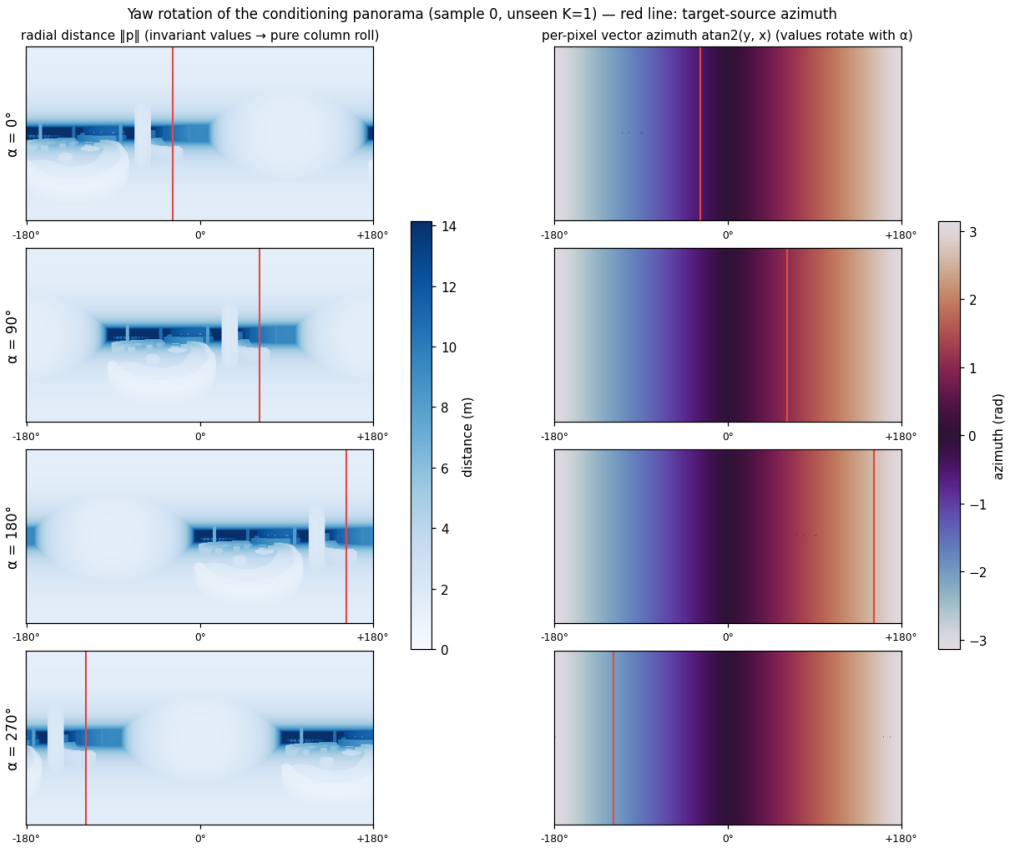

Figure 5: **Physically consistent global yaw rotation of the conditioning input.** We show one AcousticRooms example under $\alpha \in \{0°, 90°, 180°, 270°\}$. Left: radial distance, for which the scene undergoes a cyclic horizontal shift. Right: azimuth of the per-pixel 3D vectors; jointly rotating the vectors and their image locations preserves the geometric relationship between each panorama column and its associated direction. The red line indicates the target-source azimuth and rotates together with the scene.

Yaw sensitivity degrades synthesis accuracy. The heading-dependent prediction changes also reduce agreement with the ground-truth RIR. Table 7(a) shows that every nonzero rotation increases T60, C50, and EDT error and decreases retrieval recall. At $180°$, T60 error increases from 9.99% to 10.72%, C50 error from 1.047 to 1.189,dB, and EDT error from 40.11 to 46.39,ms. Across the three nonzero rotations, these changes exceed the single-evaluation variability of our baseline runs by 10–$18\times$ for T60 and 9–$17\times$ for EDT. Retrieval recall also decreases at every tested rotation and cutoff. The direct invariance gap is larger than the corresponding change in aggregate error against ground truth. For example, rotation changes EDT by about 19–20,ms between P_α and P_0, while the dataset-level EDT error increases by only 3.5–6.3,ms. Aggregate synthesis error can therefore hide part of the heading dependence because prediction changes may partially cancel when averaged against ground truth.

Overall, FLAC depends on the global heading reference despite the underlying scene being unchanged. A physically equivalent yaw reparameterization changes the predicted waveform by roughly 20% in relative ℓ_2 distance and consistently worsens synthesis accuracy. This motivates enforcing heading symmetry in the model rather than relying on it to emerge from training.

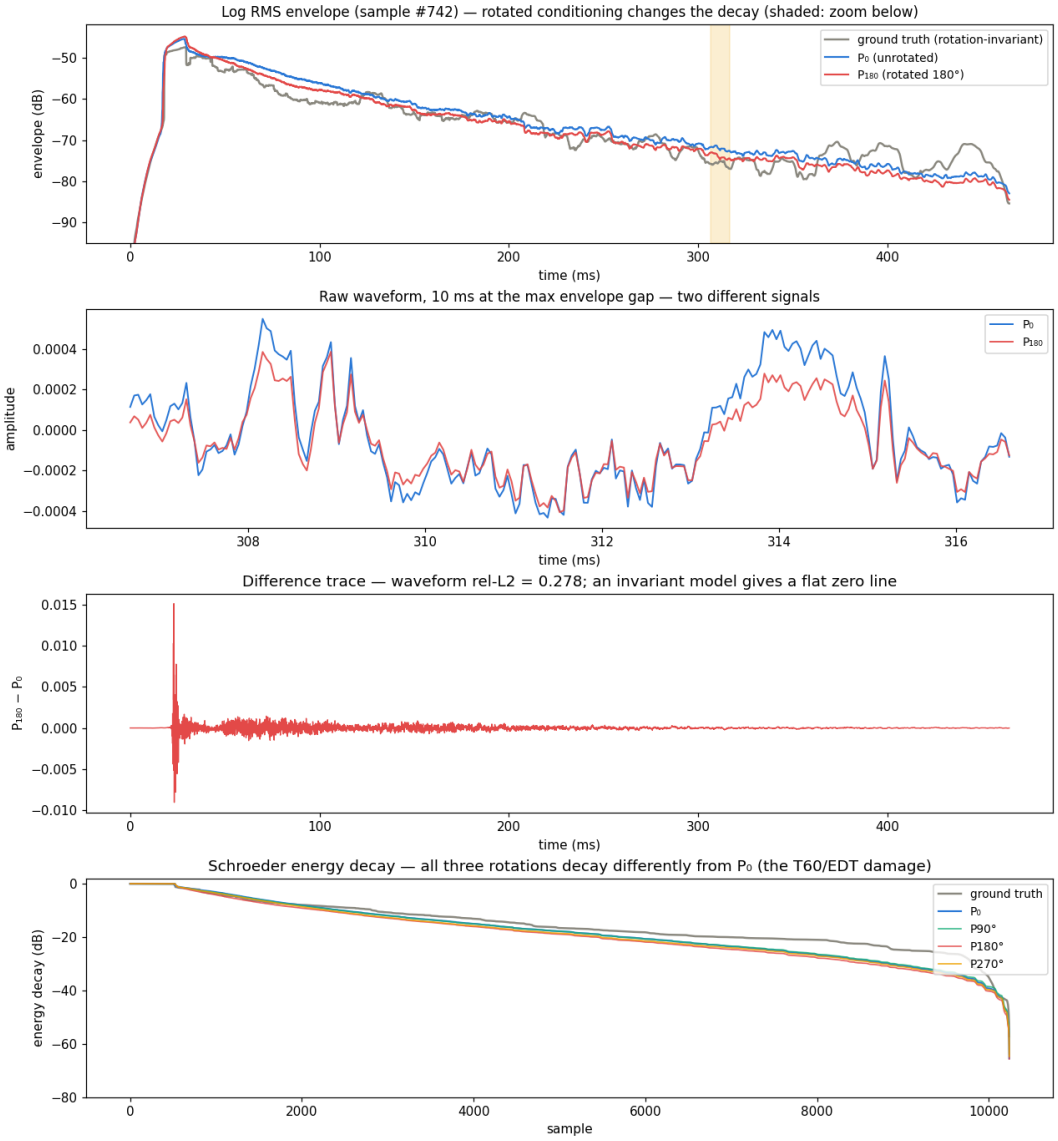

Figure 6: **Example of FLAC prediction sensitivity to the heading reference.** The ground truth, sampling noise, and physical scene are fixed; only the conditioning representation is globally rotated. The example is selected at the 90th percentile of the envelope-gap distribution rather than as the maximum-error example. From top to bottom: log-RMS decay envelopes, a 10-ms waveform segment around the largest envelope discrepancy, the waveform difference $P_{180^\circ} - P_0$, and Schroeder energy-decay curves for all three nonzero rotations. An invariant model would produce the same prediction for every rotation and hence a zero difference trace.

C.2 Encoder Equivariance and Prediction Invariance

We complement the prediction-consistency results in Figure 3 with encoder-level diagnostics. Figure 7 presents the complete comparison, examining whether geometry features transform equivariantly and whether the resulting RIR predictions remain invariant.

Joint yaw transformations. We jointly transform panoramic geometry and source–receiver poses while keeping the underlying physical acoustic configuration unchanged. For panoramic XYZ inputs, this transformation includes both a horizontal circular shift and a rotation of the horizontal coordinate channels. We evaluate patch-aligned yaw rotations, for which the spatial shift corresponds to an

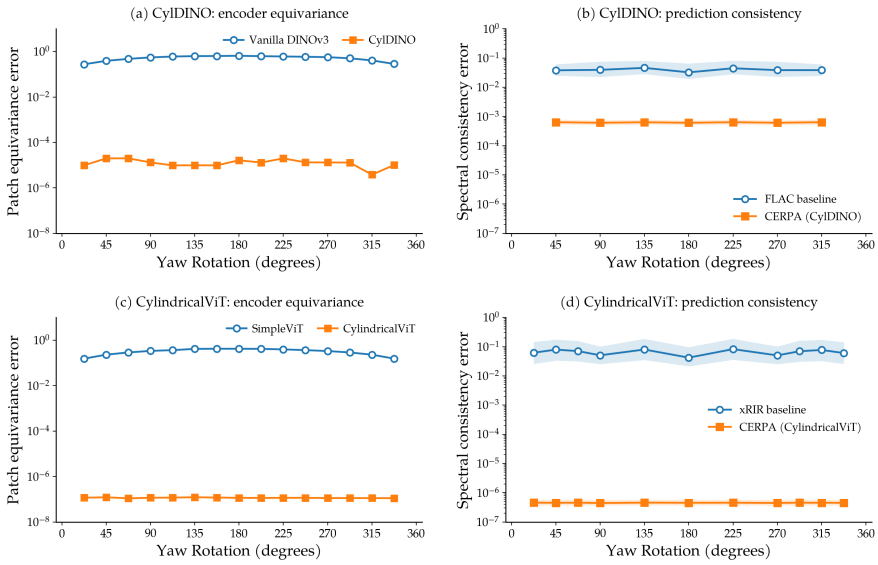

Figure 7: **Heading-equivariant representations for yaw-invariant RIR synthesis.** (a,c) CYLDINO and CYLINDRICALVIT preserve feature equivariance under yaw rotations, whereas their baseline encoders exhibit larger equivariance errors. (b,d) Both CERPA realizations produce nearly unchanged RIR predictions under patch-aligned joint yaw rotations, whereas baseline predictions vary with heading. Spectral consistency errors measure differences between predicted RIR spectra before and after rotation. Lines show median errors, and shaded bands span the 25th to 75th percentiles.

integer displacement on the patch grid. This distinguishes a change of heading reference from a physical change in the source–receiver configuration.

Encoder equivariance. Panels (a,c) measure the relative discrepancy between the original and transformed patch features after compensating for the corresponding circular shift. CYLDINO and CYLINDRICALVIT exhibit relative errors on the order of 10^{-5} and 10^{-7}, respectively, substantially below their DINOv3 and SimpleViT baselines. The small discrepancies across the evaluated angles support the intended cyclic-equivariant behavior of both geometry encoders.

RIR prediction invariance. Panels (b,d) measure the mean absolute difference between predicted log-magnitude RIR spectra before and after the joint yaw transformation on the AR unseen split. At each angle, solid curves report median errors and shaded bands span the 25th–75th percentiles. CYLDINO maintains median differences of approximately 6.3×10^{-4}, compared with 0.032–0.046 for FLAC. Similarly, CYLINDRICALVIT achieves approximately 4.5×10^{-7}, compared with 0.042–0.082 for xRIR. These consistency errors quantify sensitivity to the heading reference, rather than reconstruction error relative to ground-truth RIRs. Together, the feature- and prediction-level results support the combination of equivariant geometry processing, invariant aggregation, and invariant pose conditioning for nearly yaw-invariant acoustic synthesis under patch-aligned rotations.

C.3 EXPERIMENTAL DETAILS OF HAA FINE-TUNING

Dataset and conditioning convention. We use the *Base* configuration of the four environments in Hearing-Anything-Anywhere (HAA) (Wang et al., 2024): classroom, complex room, dampened room, and hallway. Each environment contains a loudspeaker at a fixed position and orientation, with receivers distributed throughout the room. Following the splits distributed with FLAC (Brunetto, 2026), we use 12 training receivers per room (48 in total), 560 validation queries, and 1,282 test queries. The test queries comprise 463, 198, 198, and 423 instances from the classroom, complex room, dampened room, and hallway, respectively. RIRs are processed at 22.05 kHz, and each room is represented by a 256×512 equirectangular depth panorama centred at the loudspeaker.

Unlike AcousticRooms, where contextual RIRs share a receiver, HAA contextual RIRs share a source and differ in receiver position. We interchange the source and receiver roles in the conditioning interface. For each query location, the model uses $K_{\text{ctx}} \in \{1, 8\}$ reference RIRs sampled from the same room's training set. All receiver positions are expressed relative to the fixed loudspeaker.

C.4 Experimental Details of HAA Fine-Tuning for CylDINO

Fine-tuning and evaluation protocol. We follow the released FLAC HAA fine-tuning recipe and use the same optimization settings for both variants. Each model is initialized from the EMA weights of its corresponding AcousticRooms checkpoint after 40,000 updates, with a newly initialized optimizer. Fine-tuning runs for 1,000 optimizer updates with a mini-batch size of 16 and gradient accumulation configured for four mini-batches. We use AdamW with a learning rate of 5×10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 10^{-3}. Each fine-tuning run uses a single NVIDIA RTX A6000 GPU with 48 GB of memory. CylDINO uses the variant without additional domain-specific self-supervised pretraining, with activation checkpointing enabled on the shared geometry backbone.

All reported results use the EMA weights at the fixed 1,000-update endpoint. Evaluation uses bfloat16 conditioning autocast, one sampling step, a classifier-free guidance scale of 1.0, and seeds 42–46. Following FLAC (Brunetto, 2026)'s HAA protocol, we average metrics within each room and then across rooms, excluding the dampened room from the T60 average. Each variant is trained once; reported standard deviations therefore reflect variability across evaluation seeds.

Accounting for loudspeaker directivity. Unlike the *omnidirectional* sources simulated in AcousticRooms, the loudspeakers in HAA exhibit *directivity*. Their facing direction therefore provides additional information about the acoustic relationship between the source, the room, and the receiver. To account for this property, we extend CylDINO with an explicit loudspeaker-orientation cue. This cue is treated as part of the scene and transforms jointly with its geometry, allowing the model to capture directional acoustic effects while maintaining consistency under changes of coordinate frame.

Orientation-aware cylindrical conditioning. To account for loudspeaker directivity in HAA, we condition CylDINO on the loudspeaker's facing direction. We estimate this direction exclusively from the training RIRs and represent it as a unit vector $\mathbf{f}$ in world coordinates. Broadcasting $\mathbf{f}$ over the panorama and concatenating it with the displacement field gives the six-channel input

$$\mathbf{X} = [\mathbf{x}_{\text{query}} - \mathbf{x}_{\text{surface}} \mid s\mathbf{f}] \in \mathbb{R}^{6\times 256\times 512},$$

where s controls the magnitude of the orientation input and is empirically set to 27 for HAA. The query and context conditioners use the same facing direction, paired with their respective displacement fields. We apply the cylindrical coordinate transformation to both vector fields, expressing displacement and orientation relative to each panorama column:

$$\mathcal{G}(\mathbf{X})_{u,v} = \begin{bmatrix} R_z(-\theta_u) & \mathbf{0}_{3\times 3} \\ \mathbf{0}_{3\times 3} & R_z(-\theta_u) \end{bmatrix} \mathbf{X}_{u,v}. \tag{10}$$

Here, u and v index panorama columns and rows, θ_u is the azimuth of column u, and $R_z(-\theta_u)$ rotates a vector about the vertical axis by $-\theta_u$. Under a joint yaw transformation, the facing direction rotates together with the geometry and source–receiver positions. The augmented representation therefore satisfies

$$\mathcal{G}(T_k\mathbf{X}) = \text{roll}_k(\mathcal{G}(\mathbf{X}))\,,$$

where T_k denotes a joint yaw corresponding to a horizontal shift of k panorama columns. The orientation cue thus preserves the encoder's architectural C_{32}-equivariance.

To accommodate the additional three orientation channels, we extend the patch-embedding layer with zero weights initialization. All other parameters are inherited from the 40,000-update AcousticRooms checkpoint. This initialization preserves the original conditioning outputs, allowing the model to learn how to use the orientation cue during HAA fine-tuning. The additional channels are used only for HAA adaptation, as AcousticRooms uses omnidirectional sources.

C.5 Experimental Details of HAA Fine-Tuning for CylindricalViT

We use the same HAA dataset, splits, and conditioning convention described in Appendix C.3, with $K_{\mathrm{ctx}} = 8$ for all xRIR experiments. The xRIR baseline was trained from the released checkpoint provided by xRIR paper (Liu et al., 2025).

Fine-tuning and evaluation protocol. We follow xRIR's two-stage adaptation recipe with identical optimisation settings for all variants. Each model is initialised from its AcousticRooms checkpoint after 12 epochs of pretraining. Stage 1 jointly fine-tunes on the classroom, hallway, and complex room for 1,000 full-batch epochs; stage 2 adapts the selected checkpoint to each room for 200 full-batch epochs. The respective batch sizes are 36 and 12, giving one optimiser update per epoch. Both stages use AdamW with an initial learning rate of 10^{-4}, weight decay of 10^{-4}, and exponential decay $\eta_e = 10^{-4} \times 0.1^{e/50}$. The optimiser and scheduler are reset between stages. Checkpoints are selected by validation loss, evaluated every 10 epochs in stage 1 and every two epochs in stage 2.

We run three fine-tuning seeds and report means and standard deviations across runs. Evaluation uses identical references across variants and 32 iterations of Griffin–Lim with random phase initialisation to reconstruct 9,600-sample RIRs at 22.05 kHz. Metrics are averaged within each room, with T60 omitted for the dampened room. Paired comparisons use two-way bootstrap resampling over test queries and fine-tuning seeds.

Accounting for loudspeaker directivity. HAA loudspeakers are *directional*, whereas AcousticRooms sources are *omnidirectional*. To distinguish geometrically similar configurations with different orientations relative to the loudspeaker, we provide an explicit heading cue. We infer the loudspeaker heading from directional differences in distance-compensated early RIR energy, using only the 12 training measurements in each room. In each room's original coordinate frame, this procedure selects the $-y$ direction ($\phi = -90°$). We then jointly rotate the panorama geometry and the query and reference positions by $-\phi$, aligning the estimated loudspeaker heading with the positive x-axis.

Orientation-aware cylindrical conditioning. In the heading frame, we augment the displacement field with an azimuth cue:

$$\mathbf{X} = [\mathbf{x}_{\mathrm{query}} - \mathbf{x}_{\mathrm{surface}} \mid \mathbf{a}] \in \mathbb{R}^{6\times 256\times 512},$$

where all positions are normalized following xRIR's coordinate convention, and the query position is broadcast over the panorama. The azimuth cue is defined as

$$\mathbf{a}_{u,v} = \begin{bmatrix} \cos\theta_u \\ \sin\theta_u \\ 0 \end{bmatrix},$$

where θ_u is the azimuth of panorama column u in the heading frame. This cue encodes each column's horizontal direction relative to the estimated loudspeaker heading. Before patch embedding, we apply the cylindrical transformation to the displacement channels while leaving the azimuth cue unchanged:

$$\mathbf{Z}_{u,v} = \begin{bmatrix} R_z(-\theta_u) & \mathbf{0}_{3\times 3} \\ \mathbf{0}_{3\times 3} & I_3 \end{bmatrix} \mathbf{X}_{u,v}.$$

The shared geometry encoder uses the same cue with the respective displacement fields for all conditioning locations. Heading alignment produces consistent inputs under joint rotations of the scene and heading reference corresponding to integer panorama-column shifts.

In implementation, the identically zero component is omitted before patch embedding, so the embedding layer is expanded from three to five input channels, while the remaining CylindricalViT layers are unchanged. This architecture is pretrained from scratch on AcousticRooms for 12 epochs with a fixed heading of $0°$, then fine-tuned on HAA.

Table 6: **Model architecture and training configuration of CylindricalViT.** The model is trained from scratch on the AcousticRooms training split for unseen-room evaluation using a single NVIDIA RTX A6000 GPU. Reported comparisons use the final checkpoint after 12 epochs, corresponding to 55,572 optimization steps.

Category	Parameter	Configuration
Architecture	Geometry encoder	CylindricalViT
	Model initialization	Random initialization
	Backbone sharing	Shared across target-source, receiver, and reference-source geometry encoding
	Transformer depth / hidden dimension / heads	12 / 512 / 8
	Attention head dimension / MLP dimension	64 / 512
	Acoustic encoder	ResNet-18 on reference log-magnitude spectrograms; 512-dimensional output
	Geometry and fusion feature dimension	256
	Prediction head	Time-dependent weighted combination of aligned reference log-magnitude spectrograms
Input	Geometry representation	XYZ displacement field, $(\mathbf{x}_{\mathrm{query}} - \mathbf{x}_{\mathrm{surface}})/5$
	Input dimensions / patch size	$3 \times 256 \times 512$ / 16×32 pixels
	Patch grid / number of patch tokens	16×16 / 256
	RIR channels / sampling rate	Single-channel / 22,050 Hz
	Waveform length	9,600 samples (≈ 0.435 s)
	STFT configuration	FFT size 124, Hann window length 62, hop length 31; 63×310 magnitude spectrogram
Geometry encoding	Cylindrical XYZ gauge	Per-column rotation of horizontal XYZ components into the local cylindrical basis
	Absolute positional encoding	Elevation-only sinusoidal encoding, shared across azimuth columns
	Relative positional encoding	Learned per-head attention bias indexed by relative elevation and circular azimuth offsets
Training	Hardware	$1\times$ NVIDIA RTX A6000 GPU (48 GB)
	Training data	296,334 RIRs from 243 training rooms
	Optimizer	AdamW; initial learning rate 10^{-3}, $\beta = (0.9, 0.999)$, weight decay 10^{-4}
	Training budget	12 epochs; 55,572 optimization steps
	Per-GPU batch / gradient accumulation	32 / 2 mini-batches per update
	Effective batch size	64, except for the final partial update in each epoch
	Learning-rate schedule	Multiplied by $0.1^{1/3} \approx 0.46416$ after each epoch, giving a tenfold reduction every three epochs
	Training objective	STFT magnitude L_1 loss plus $0.01 \times L_1$ loss on normalized energy-decay curves in dB
	Acoustic context	$K_{\mathrm{ctx}} = 8$
	Checkpoint evaluation / final selection	Evaluated after every epoch / final epoch-12 checkpoint
Evaluation	Test split	6,337 queries from 17 held-out rooms
	Acoustic context size	$K_{\mathrm{ctx}} \in \{1, 8\}$
	Batch size	16
	Evaluation seeds	42–46 for reference selection and phase initialization; results reported as mean $\pm$ standard deviation

Table 7: **Effect of changing the heading reference on the frozen FLAC model.** (a) RIR synthesis accuracy with respect to the same ground-truth RIR. Lower is better for T60, C50, EDT, and FD; higher is better for retrieval recall. (b) Direct difference between the rotated prediction P_α and the unrotated prediction P_0, without using ground truth. A yaw-invariant model would have zero gap in all columns of (b).

(a) Accuracy against ground truth

Rotation	T60 (%) ↓	C50 (dB) ↓	EDT (ms) ↓	FD ↓	R@1 ↑	R@5 ↑	R@10 ↑
0°	9.99	1.047	40.11	0.3027	6.71	19.00	26.89
90°	10.38	1.074	43.58	0.3088	6.30	17.47	25.09
180°	10.72	1.189	46.39	0.3032	6.38	17.85	25.14
270°	10.44	1.126	44.07	0.3088	6.11	17.80	25.33

(b) Direct invariance gap between P_α and P_0

α	Waveform rel-ℓ_2	Mean abs. diff.	T60 gap (%)	C50 gap (dB)	EDT gap (ms)
0°	0.000	0	0.00	0.000	0.00
90°	0.221	1.38×10^{-4}	3.34	0.563	18.65
180°	0.193	1.21×10^{-4}	3.33	0.550	20.11
270°	0.214	1.35×10^{-4}	3.41	0.600	19.98