# Few-shot Acoustic Synthesis with Multimodal Flow Matching

Amandine Brunetto
Mines Paris - PSL University
https://amandinebtto.github.io/

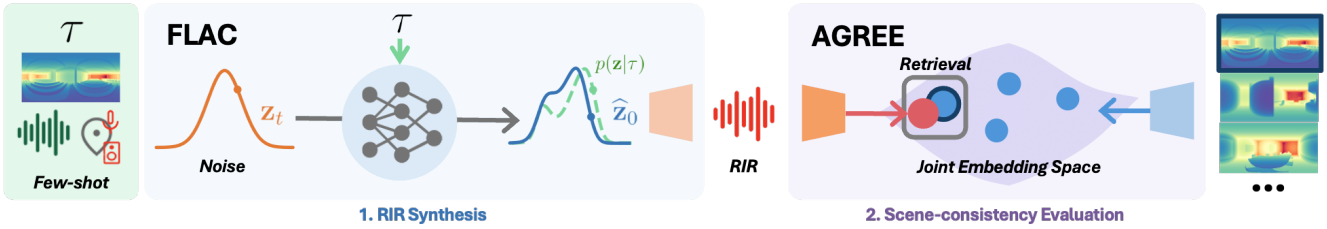


Figure 1. **Few-shot flow-matching acoustic synthesis (FLAC) and scene-consistency evaluation:** Given a few-shot multimodal context $\tau$, including a depth map, an acoustic observation, and sensor poses, FLAC uses a diffusion transformer trained with flow matching to generate room impulse responses (RIRs) in novel rooms. Unlike prior deterministic approaches, FLAC models the distribution of plausible RIRs under sparse scene context, capturing acoustic uncertainty. Even with one shot, FLAC outperforms 8-shot state-of-the-art methods. To assess generation quality, we introduce AGREE, a CLIP-style audio-geometry embedding that aligns both modalities in a shared latent space, enabling scene-consistency evaluation through retrieval and distributional metrics.

## Abstract

*Generating audio that is acoustically consistent with a scene is essential for immersive virtual environments. Recent neural acoustic field methods enable spatially continuous sound rendering but remain scene-specific, requiring dense audio measurements and costly training for each environment. Few-shot approaches improve scalability across rooms but still rely on multiple recordings and, being deterministic, fail to capture the inherent uncertainty of scene acoustics under sparse context. We introduce flow-matching acoustic generation (FLAC), a probabilistic method for few-shot acoustic synthesis that models the distribution of plausible room impulse responses (RIRs) given minimal scene context. FLAC leverages a diffusion transformer trained with a flow-matching objective to generate RIRs at arbitrary positions in novel scenes, conditioned on spatial, geometric, and acoustic cues. FLAC outperforms state-of-the-art eight-shot baselines with one-shot on both the AcousticRooms and Hearing Anything Anywhere datasets. To complement standard perceptual metrics, we further introduce AGREE, a joint acoustic–geometry embedding, enabling geometry-consistent evaluation of generated RIRs through retrieval and distributional metrics. This work is the first to apply generative flow matching to explicit RIR synthesis, establishing a new direction for robust and data-efficient acoustic synthesis. Project page: https://amandinebtto.github.io/FLAC/*

## 1. Introduction

Every room shapes the way we hear: a lecture hall amplifies a speaker's voice, while a cathedral envelops sound in lingering reverberation. Reproducing these rich auditory experiences is essential for creating virtual, immersive environments, where users expect sound to reflect the space.

The acoustic properties of a room are encapsulated by Room Impulse Responses (RIRs), which describe the sound propagation between source-receiver pairs. RIRs allows for auralization, *i.e.*, transferring a room's acoustic signature onto any sound. However, accurately modeling RIRs is challenging because they depend on complex interactions between geometry, materials, and source-listener positions.

Recently, neural acoustic fields [2, 4, 10, 41, 51, 74, 77] have enabled spatially continuous RIRs rendering in a scene. However, they must be trained for each environment using extensive RIR recordings. More scalable solutions require models that can generate RIRs in novel rooms, with minimal data and without retraining.

A handful of works have explored few-shot acoustic synthesis [34, 49, 54]. These methods generate RIRs in novel environments using only a sparse set of information (*e.g.*, depth maps, RGB images, sensor poses, and 8 to 20 RIR recordings) without scene-specific retraining. With limited knowledge about a new scene's characteristics, there is no single, deterministic possible RIR: few-shot generalization is an inherently ambiguous problem. Yet, existing few-shot methods overlook this uncertainty, producing only a unique

deterministic prediction.

To address this, we propose FLAC, a conditional generative model for few-shot acoustic synthesis based on flow matching [44]. This framework extends diffusion models [33, 71] with increased performance and versatility and has demonstrated strong performance in audio [36, 39, 46] and images [18] generation. FLAC is, to the best of our knowledge, the first application of generative flow matching to explicit RIR synthesis. Rather than learning a deterministic mapping, our model estimates a distribution of plausible RIRs given sparse scene context, explicitly capturing the uncertainty inherent in few-shot scenarios. We condition the generation on multimodal context, including scene geometry around the receiver, sensor poses, and a minimal set of RIR recordings. By formulating few-shot acoustic synthesis as a conditional generative task, we enable scene-consistent sound generation in novel environments even from only one audio measurement.

To assess generation quality, we complement traditionally used perceptual metrics by introducing a set of scene consistency metrics that ensure the predicted RIR matches the scene's geometry. To this end, we introduce AGREE (Acoustic-GeometRy EmbEdding), a CLIP-style [62] dual-encoder network that aligns RIRs and scene geometry in a shared latent space. This alignment enables zero-shot audio and geometry retrieval. We leverage this shared space to provide a geometry-consistent evaluation framework through both retrieval-based scores and distributional metrics.

We evaluate FLAC on the large-scale synthetic AcousticRooms [49] dataset. It achieves state-of-the-art RIR synthesis performance, demonstrating generalization across novel source-receiver positions within known rooms, as well as in entirely unseen environments. We also validate our model's real-world capabilities through sim-to-real transfer on the Hearing-Anything-Anywhere [77] dataset. On both datasets, FLAC outperforms current state-of-the-art methods based on 8 audio recordings with a single one.

In summary, our main contributions are as follows:

- We propose FLAC the first conditional generative model for few-shot RIR synthesis based on flow matching. This approach accounts for the inherent uncertainty of acoustics given sparse scene context, leading to more robust predictions.
- Our approach sets a new state-of-the-art on the AcousticRooms and Hearing-Anything-Anywhere datasets, generalizing to both novel source-listener pairs and environments. FLAC outperforms prior work with $8\times$ fewer RIR recordings.
- We introduce AGREE, a joint acoustic-geometry embedding space, and propose new scene-consistency metrics that evaluate how well predicted RIRs align with the scene geometry.

## 2. Related Work

**Audio-visual learning.** Audio-visual learning enhances both acoustic and vision-related tasks, including audio spatialization [22, 25, 38, 57, 76, 81], de-reverberation [9, 13], RIR prediction [7, 41, 42, 49, 54, 69, 70], depth estimation [3, 14, 59, 80, 82], navigation [6, 11, 21, 23, 24, 79], floorplan reconstruction [55, 61], and pose estimation [11]. FLAC extends this line of work by leveraging depth information for scene-consistent RIR generation.

**Neural acoustic fields.** Neural acoustic fields render RIRs at novel poses by implicitly learning a mapping from spatial coordinates to the room's acoustic field. Some approaches incorporate physical acoustic models [74, 77], others infer local geometry [51], exploit vision cues [10, 12, 41, 42] or use NeRF [56] and Gaussian splatting-based [37] representations [2, 4]. However, these methods remain scene-specific, requiring dense recordings and retraining for each new environment.

**Few-shot acoustic synthesis.** Few-shot methods generalize across scenes using sparse observations. FewShotRIR [54] uses 20 RGB, depth and binaural audio inputs. MAGIC [34] adds semantics by extracting features with a segmentation-pretrained U-Net [66]. More recently, xRIR [49] reduces inputs to eight audio recording and a panoramic depth map, and introduces the AcousticRooms dataset specifically designed for cross-room synthesis. All prior methods treat few-shot RIR prediction as a deterministic mapping, overlooking the ambiguity of the task. By using generative flow matching, FLAC captures the distribution of plausible RIRs given sparse context, improving generalization to new scenes even with one-shot.

**Audio diffusion and flow matching.** Diffusion-based models have advanced text-to-audio generation across speech, music, and general sound [19, 20, 26, 35, 45, 46, 53]. Flow matching further improves synthesis efficiency [28, 36, 39]. [43] recently achieved speech binauralization via flow matching. Building on these advances, we adapt generative flow matching to RIR synthesis conditioned on few-shot scene context.

**Joint embedding models across modalities.** Joint embedding models align data from different modalities in a shared representation space. CLIP [62] pioneered this for image-text, later extended to audio-visual [29, 52, 58, 64], audio-text [17, 29], and audio with diverse sensory modalities [27], enabling zero-shot cross-modal retrieval. Standard audio embeddings cannot be applied directly to RIRs, which differ substantially. We introduce AGREE, a joint embedding space for RIRs and scene geometry, allowing acoustic-geometry consistency evaluation.

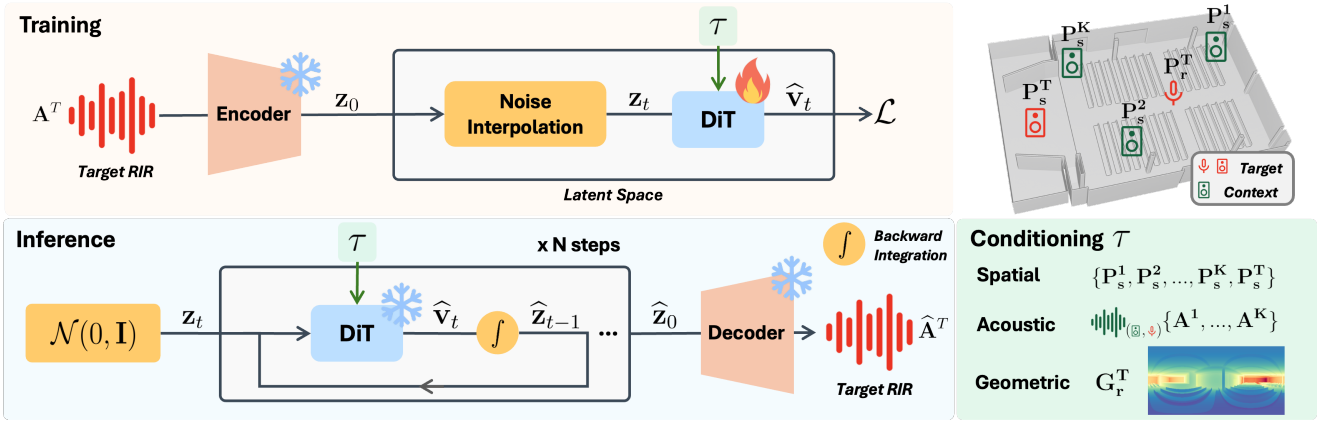


Figure 2. **Training and inference pipelines of FLAC:** During training, a pre-trained VAE encodes ground-truth RIRs into latents $\mathbf{z}_0$. Latents are linearly interpolated with noise to form $\mathbf{z}_t$. A DiT is trained to predict the velocity $\hat{\mathbf{v}}_t$ that transports $\mathbf{z}_t$ toward the original data distribution. At inference, RIRs are generated from random noise, guided by the few-shot spatial, geometric and acoustic context.

## 3. Method

FLAC is a conditional latent generative model [65] trained with flow matching [1, 44] (Sec. 3.1) to synthesize RIRs from few-shot scene information. It comprises: (i) a variational autoencoder (Sec. 3.2), (ii) a multimodal conditioner (Sec. 3.3), and (iii) a diffusion transformer (Sec. 3.4). Fig. 2 provides an overview of the method.

### 3.1. Latent Flow Matching

**Ambiguity in few-shot synthesis.** Estimating RIRs across diverse environments and sensor poses is challenging, as they depend on many intertwined factors. With limited scene information, multiple RIRs can be equally plausible for the same source-receiver configuration. For instance, even with precise geometry knowledge, missing material properties introduces ambiguity: whether the floor is carpeted or wooden alters the acoustics.

We address the inherently ambiguous problem of few-shot RIR synthesis: Our goal is to predict monaural, omnidirectional RIRs at arbitrary source-receiver pairs in unseen environments, given minimal scene context. By using a stochastic generative model, we aim to capture the uncertainty inherent to RIR prediction under sparse observations.

**Training.** We train FLAC using the rectified flow matching formulation [47, 48], which linearly interpolates data and noise. This approach straightens the transport paths between distributions, reducing the number of integration steps at inference.

The goal is to capture the relationship between a RIR and its spatial, geometric, and acoustic context. To this end, we sample target RIRs with their associated context $(A^T, \boldsymbol{\tau})$ from the dataset. Each RIR is encoded into a latent representation $\mathbf{z}_0$, which is linearly interpolated with Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ to produce a noisy latent $\mathbf{z}_t$:

$$\mathbf{z}_t = (1-t)\,\mathbf{z}_0 + t\,\boldsymbol{\epsilon}, \tag{1}$$

where the timestep $t \in [0,1]$ controls the noise level. Timesteps are sampled by drawing $\alpha \sim \mathcal{N}(-1.2, 2^2)$ and mapping it to $t$ using a sigmoid:

$$t = \sigma(-\boldsymbol{\alpha}) = \frac{1}{1+e^{\boldsymbol{\alpha}}}. \tag{2}$$

This schedule emphasizes on moderately noisy latents ($t \approx 0.7{-}0.8$), which we found to improve performance. Comparisons of noise sampling strategies are provided in Appendix E.4.

The model $u(\mathbf{z}_t, t, \boldsymbol{\tau})$ is trained to predict the velocity field $\mathbf{v}_t$

$$\mathbf{v}_t = \frac{d\mathbf{z}_t}{dt} = \boldsymbol{\epsilon} - \mathbf{z}_0, \tag{3}$$

using the following objective:

$$\mathcal{L}_{\text{RFM}} = \mathbb{E}_{\mathbf{z}_0,\boldsymbol{\epsilon},t,\boldsymbol{\tau}}\Big[\,\|\,u(\mathbf{z}_t,t,\boldsymbol{\tau}) - \mathbf{v}_t\|^2\Big]. \tag{4}$$

**Inference.** We employ classifier-free guidance [32], allowing the model to learn both conditional and unconditional distributions by randomly dropping the conditioning during training.

At inference, the guided velocity prediction is given by

$$\hat{u}(\mathbf{z}_t,t,\boldsymbol{\tau}) = u(\mathbf{z}_t,t,\varnothing) + \omega\,\big[u(\mathbf{z}_t,t,\boldsymbol{\tau}) - u(\mathbf{z}_t,t,\varnothing)\big], \tag{5}$$

where $\omega > 0$ controls the conditioning strength, and $u(\mathbf{z}_t,t,\varnothing)$ denotes the unconditional prediction.

RIRs are generated by solving the ordinary differential equation (ODE) backward, starting from Gaussian noise $\boldsymbol{\epsilon}$ and integrating the velocity field from $t=1$ to $t=0$:

$$\mathbf{z}_{t-dt} = \mathbf{z}_t + \hat{u}(\mathbf{z}_t,t,\boldsymbol{\tau})\,dt. \tag{6}$$

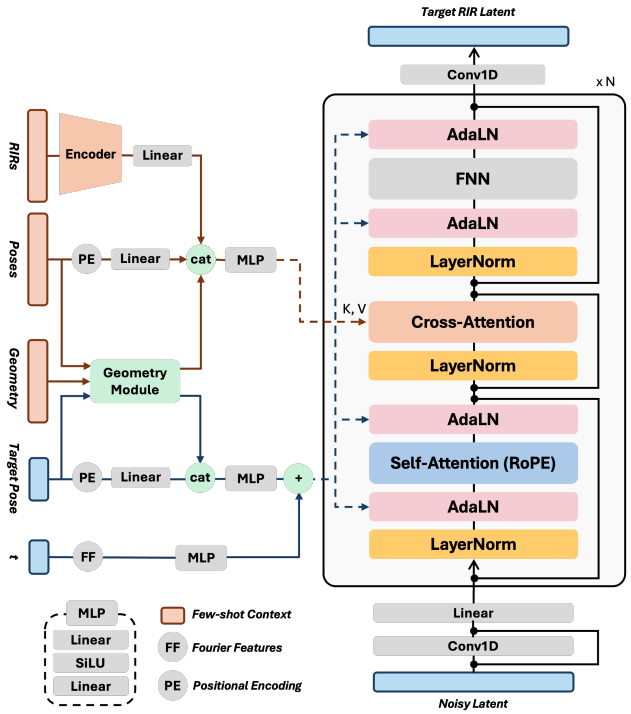


Figure 3. **FLAC diffusion transformer:** The noise timestep $t$ and the target RIR pose are injected via AdaLN. Acoustic, spatial and geometric context are provided through cross-attention.

### 3.2. VAE

We train a variational autoencoder (VAE) to compress RIR waveforms into $\mathbf{z}_0$. The encoder consists of four convolutional blocks, each performing downsampling and channel expansion with strided convolutions. Before each downsampling block, we apply ResNet-style layers with dilated convolutions and Snake activations [83]. The bottleneck has a latent feature dimension of 32, and the decoder mirrors the encoder. All convolutions are weight-normalized [67] and the output passes through a $\tanh$ activation to match the RIR amplitude range.

We found pre-trained audio embeddings unsuitable for latent flow matching. Obtaining a compact RIR representation is challenging as it must preserve precise temporal and spectral structure. To achieve this, we train the VAE with complementary objectives: a multiresolution STFT loss $\mathcal{L}_{\text{MR}}$ [72, 78] combining spectral convergence, spectral and energy decay terms; an adversarial hinge loss $\mathcal{L}_{\text{adv}}$; a feature-matching loss $\mathcal{L}_{\text{feat}}$ using Encodec [15] multi-scale STFT discriminator; and a KL divergence loss $\mathcal{L}_{\text{KL}}$ to regularize the latent space. The final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{MR}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{KL}} \tag{7}$$

Details on the implementation, individual loss terms, and hyperparameters are provided in Appendix A.

### 3.3. Multimodal Conditioning

FLAC generates RIRs at a target source-receiver pair $(P_s^T, P_r^T)$ based on multimodal scene context $\boldsymbol{\tau}$:

- **Acoustic:** RIRs measured at the target receiver $P_r^T$ from $K$ different source positions, $\mathbf{A} = \{A^1, \ldots, A^K\}$, capturing key room acoustic properties.
- **Spatial:** Corresponding source positions $\mathbf{S} = \{P_s^1, \ldots, P_s^K\}$ and the target source position $P_s^T$.
- **Geometric:** A panoramic depth map $\mathbf{G}_r^T$ captured at the target receiver pose $P_r^T$, describing local room structure and surfaces.

Below, we detail how each modality is processed.

**Acoustic.** Similar to [49, 54] each of the $K$ context RIRs is transformed into a magnitude spectrogram and encoded with a ResNet-18 backbone [30], trained jointly with the rest of the model. The encoder outputs a 512-dimensional embedding per RIR, capturing key acoustic properties.

**Spatial.** Since the receiver is shared between the context and target RIRs, we express all source poses in the receiver's local coordinate frame and omit $P_r^T$ (the origin). The resulting 3D coordinates are encoded with sinusoidal positional embeddings and projected into a high-dimensional feature space through a linear layer.

**Geometric.** We condition on the geometry surrounding the receiver to capture the location and shape of nearby surfaces. A panoramic depth map captured at $P_r^T$ is converted into an image containing 3D coordinates via equirectangular projection. Following [49], we compute reflection maps by subtracting each source position (target and $K$ context) expressed in the receiver's frame from these 3D coordinates. DINOv3 [68] Vision Transformer (ViT) [16] S/16 is fine-tuned to encode the reflection maps into compact features capturing geometric structure and spatial relationships. An overview of the geometry module is given in Appendix B.2.

### 3.4. Diffusion Transformer

Inspired by recent advances in image and audio generation [19, 20, 36, 40, 60], we parameterize the velocity field $\widehat{\mathbf{v}}_t$ using a diffusion transformer (DiT) illustrated in Fig. 3. It consists of a multi-layer Transformer architecture. A 1D convolution followed by a linear layer maps between the VAE latent space and the transformer embedding dimension. Each transformer block follows a fixed sequence: self-attention with Rotary Positional Embedding (RoPE) [73], followed by cross-attention over conditioning tokens and a feedforward network (FNN), with residual connections applied inside each sub-layer. We compute $d$-dimensional Fourier features of the noise timestep $t$. The global conditioning, containing the target pose and $t$, is incorporated via Adaptive Layer Norm (AdaLN), where learned scale, shift and gating parameters modulate both self-attention and

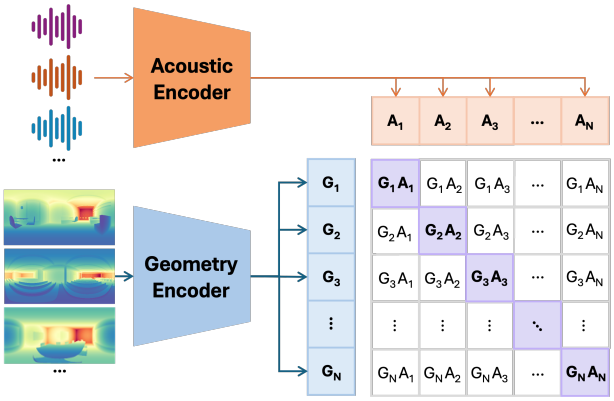


Figure 4. **AGREE contrastive framework:** Audio and geometry inputs are encoded into a shared latent space, where a contrastive objective maximizes similarity for matching pairs (diagonal entries) and minimizes it for mismatched ones.

feedforward layers. Acoustic, spatial and geometric context are incorporated through cross-attention. Finally, the model consists of 12 transformer blocks with 8 heads and a hidden width of 256. We train it with a learning rate of $5 \times 10^{-5}$, AdamW optimizer [50] and a batch size of 64 on a single H100 GPU. We use an Exponential Moving Average (EMA) of the model weights during training and BF16 precision. FLAC number of parameters and inference time are reported in the Appendix H.

## 4. AGREE: Acoustic–Geometry Embedding

We introduce AGREE (Acoustic-GeometRy EmbEding), a CLIP-style [62] multimodal embedding that aligns room acoustics and geometry (see Fig. 4). The audio encoder is fine-tuned from the pre-trained VAE encoder used in FLAC (see Sec. 3.2). The geometry encoder is DINOv3 ViT-S/16 fine-tuned on panoramic depth maps captured at receiver positions. Following FLAC's geometry pipeline, depth values are projected into 3D coordinates and source positions, expressed in the receiver frame, are subtracted. Each encoder is followed by a linear projection, and both are trained jointly with a contrastive objective to align acoustic and geometric representations. The resulting embedding space captures spatial-acoustic consistency, enabling zero-shot cross-modal retrieval and geometry-aware evaluation. AGREE details are provided in Appendix C.

## 5. Experiments

### 5.1. Datasets

**AcousticRooms.** We use the AcousticRooms (AR) dataset [49], a large-scale simulated dataset of monaural RIRs paired with equirectangular panoramic depth maps. It spans 260 rooms across 10 categories with diverse geometries, sizes, and materials, totaling over 300k simulated RIRs at 22,050 Hz. Generated with Treble Technology's wave-based simulation, it provides high simulation accuracy beyond geometric or ray-tracing methods used in [5, 8, 75]. Following [49], we split the dataset into 243 seen and 17 unseen rooms to evaluate both in-room prediction and generalization to new scenes. The unseen test set contains 5,244 instances. A subset of the seen-room instances is used for evaluation, it contains 6,217 instances across 131 rooms. In all our experiments, our VAE is pretrained on this dataset.

**Hearing-Anything-Anywhere.** To evaluate generalization to real-world environments, we use the Hearing-Anything-Anywhere (HAA) dataset [77]. It provides monaural RIRs recorded in four rooms, each with a fixed source and multiple receiver positions. This setup is the inverse of AcousticRooms, where the receiver is fixed and the source varies. However, for single-channel RIRs, interchanging source and receiver is equivalent due to the symmetry of the wave equation [49]. All RIRs are sampled to 22,050 Hz. Panoramic depth maps at each source pose are derived from room meshes reconstructed using wall and surface annotations. Appendix D.1 gives datasets details.

### 5.2. Metrics

**Perceptual metrics.** We assess the perceptual quality of generated RIRs using standard acoustic metrics [4, 41, 49, 54, 74] that correlate with human auditory perception. We report the relative T60 error, normalized by the ground-truth. T60 measures the reverberation time *i.e.*, the duration for sound energy to decay by 60 dB. We also compute the clarity error based on C50, the ratio of early-to-late energy, indicative of speech intelligibility and acoustic clarity. Finally, we evaluate the Early Decay Time (EDT) error, which capture early reflection characteristics by measuring the time for an initial 5 dB energy decay.

**Scene-consistency metrics.** We introduce metrics based on the AGREE embedding space to evaluate how well generated RIRs reflects the spatial characteristics of the environment. We compute audio-to-audio recall (R@1/5/10), quantifying how closely generated and ground-truth RIRs align in this geometry-aware space. To capture overall realism, we compute the Fréchet distance $\text{FD}_G$, between the distribution of generated and real audio embeddings in AGREE space, analogous to the FID [31] used in image generation. For reference, AGREE zero-shot retrieval results on the unseen AcousticRooms set are summarized in Tab. 2. To maximize retrieval performance when evaluating RIR synthesis methods, we also train AGREE on the entire AR dataset. Further details on the scene-consistency metrics can be found in Appendix D.3.

Table 1. **Performance on unseen AcousticRooms scenes:** Results are shown for $K \in \{8, 1, ✗\}$ reference RIRs. For FLAC, we report mean and standard deviation over 5 generations. FLAC outperforms all baselines even in the one-shot setting. ✗ denotes ablations with either geometry (G) or audio conditioning removed.

| Method | K | G | T60 (%) ↓ | C50 (dB) ↓ | EDT (ms) ↓ | R@1 (%) ↑ | R@5 (%) ↑ | R@10 (%) ↑ | $FD_G$ ↓ |
|---|---|---|---|---|---|---|---|---|---|
| Random Across Rooms | ✗ | ✗ | 44.73 | 7.676 | 306.29 | 0.02 | 0.06 | 0.32 | 0.111 |
| Random Same Room | ✗ | ✗ | **17.36** | 5.490 | 168.17 | 0.25 | 1.09 | 2.16 | **0.001** |
| FLAC✗ | ✗ | ✓ | $23.41_{\pm 0.02}$ | $\mathbf{2.554}_{\pm 0.002}$ | $\mathbf{109.75}_{\pm 0.09}$ | $\mathbf{5.12}_{\pm 0.12}$ | $\mathbf{16.47}_{\pm 0.14}$ | $\mathbf{23.12}_{\pm 0.14}$ | 0.337 |
| Nearest Neighbor | 1 | ✗ | 15.22 | 5.212 | 157.94 | 0.00 | 2.26 | 4.56 | **0.001** |
| Fast-RIR | 1 | ✓ | 18.97 | 3.257 | 121.21 | 0.17 | 0.66 | 1.64 | 0.456 |
| xRIR | 1 | ✓ | 14.47 | 1.961 | 74.45 | 0.28 | 1.36 | 2.59 | 0.263 |
| **FLAC** | 1 | ✓ | $\mathbf{9.95}_{\pm 0.05}$ | $\mathbf{1.046}_{\pm 0.002}$ | $\mathbf{40.04}_{\pm 0.22}$ | $\mathbf{6.80}_{\pm 0.11}$ | $\mathbf{18.92}_{\pm 0.10}$ | $\mathbf{26.87}_{\pm 0.19}$ | 0.303 |
| Linear Interpolation | 8 | ✗ | 14.45 | 3.503 | 114.27 | 0.41 | 2.30 | 4.02 | 0.401 |
| Nearest Neighbor | 8 | ✗ | 10.91 | 2.792 | 90.08 | 0.00 | 10.26 | 17.28 | **0.003** |
| FLAC✗ | 8 | ✗ | $12.07_{\pm 0.01}$ | $4.296_{\pm 0.001}$ | $140.04_{\pm 0.04}$ | $0.09_{\pm 0.01}$ | $0.58_{\pm 0.06}$ | $1.06_{\pm 0.04}$ | 0.663 |
| Fast-RIR | 8 | ✓ | 17.71 | 3.253 | 121.21 | 0.24 | 0.99 | 1.88 | 0.465 |
| xRIR | 8 | ✓ | 9.98 | 1.354 | 49.40 | 0.54 | 2.00 | 3.38 | 0.307 |
| **FLAC** | 8 | ✓ | $\mathbf{8.60}_{\pm 0.01}$ | $\mathbf{0.970}_{\pm 0.002}$ | $\mathbf{37.13}_{\pm 0.02}$ | $\mathbf{6.99}_{\pm 0.13}$ | $\mathbf{19.38}_{\pm 0.15}$ | $\mathbf{27.21}_{\pm 0.17}$ | 0.305 |

Table 2. **Zero-shot cross-modal retrieval on the unseen AcousticRooms set:** We report acoustic-to-geometry (A2G) and geometry-to-acoustic (G2A) recall at 1, 5 and 10. † indicates training on the full dataset for benchmarking few-shot methods.

| Method | A2G | | | G2A | | |
|---|---|---|---|---|---|---|
| | R@1↑ | R@5↑ | R@10↑ | R@1↑ | R@5↑ | R@10↑ |
| AGREE | 59.78 | 83.53 | 89.35 | 59.10 | 85.56 | 91.04 |
| AGREE† | 85.37 | 99.70 | 99.98 | 84.38 | 99.53 | 99.97 |

## 5.3. Baselines

We compare FLAC against several baselines:

- Random Across Rooms: randomly samples a RIR from the entire dataset.
- Random Same Room: randomly selects a RIR from the same room.
- Linear Interpolation: linearly interpolates $K$ reference RIRs based on their distances to the target source.
- Nearest Neighbor (KNN): chooses the RIR closest in distance to the target source among the $K$ references.
- Fast-RIR [63]: generates RIRs with a GAN conditioned on T60 and scene size estimated from $K$ RIRs and depth.
- xRIR [49]: combines acoustic and geometric features to weight $K$ reference RIRs.

## 5.4. Inference parameters

In all experiments, we use a guidance scale of 1 and perform generation in a single inference step as it achieves the best results on the perceptual metrics (T60, C50, EDT). These metrics mainly capture global acoustic properties, such as energy decay and clarity, but are insensitive to fine-grained details or sample diversity. Thus, additional steps offer no benefit. However, as shown in Fig. 5, increasing the guidance weight or the number of steps improves $FD_G$.

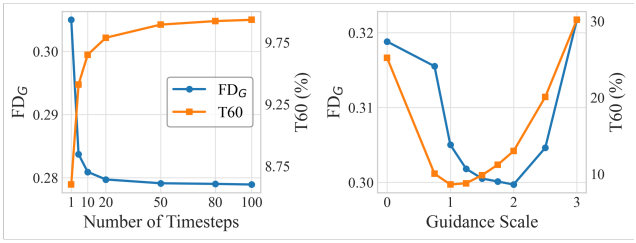


Figure 5. **Impact of classifier-free guidance and inference steps:** Evolution of T60 and $FD_G$ as a function of the guidance scale $\omega$ and the number of timesteps.

## 5.5. Results

**8-shot generation in novel environments.** Quantitative results on unseen scenes with $K$=8 are reported in Tab. 1. FLAC consistently outperforms xRIR across all metrics, reducing errors by 13.8% T60, 28.3% C50, and 24.9% EDT, and achieving higher audio-to-audio recall, indicating more geometry-consistent acoustic synthesis. For $FD_G$, FLAC slightly surpasses xRIR, reflecting improved distributional realism. Increasing the number of inference steps or the classifier-free guidance weight further improves FLAC $FD_G$; for example, 20 steps reduce it to 0.280 (see Fig. 5). KNN always achieves lower $FD_G$ as it simply returns a reference RIR, which is already drawn from the true distribution. For seen rooms, detailed results are provided in Appendix E.1: FLAC reduces errors by 23.9%, 29.8%, and 24.8% for T60, C50, and EDT, respectively. These results demonstrate that FLAC not only improves RIR estimation at new positions within seen spaces but also generalizes more effectively to new environments.

**Robustness under limited observations.** We evaluate methods robustness with fewer context RIRs, simulating scenarios with limited recordings. For FLAC and xRIR, models trained with $K$=8 and tested with fewer references.

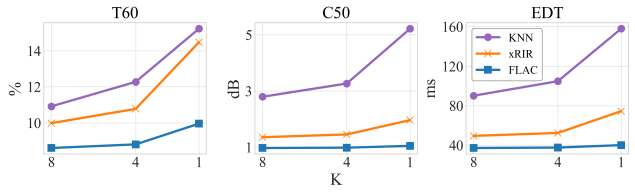


Figure 6. **Robustness to limited context RIRs in novel scenes:** Performance as the number of reference RIRs ($K$) decreases for KNN, xRIR, and FLAC. FLAC remains the most stable and outperforms state-of-the-art methods with $K$=8 even in one-shot.

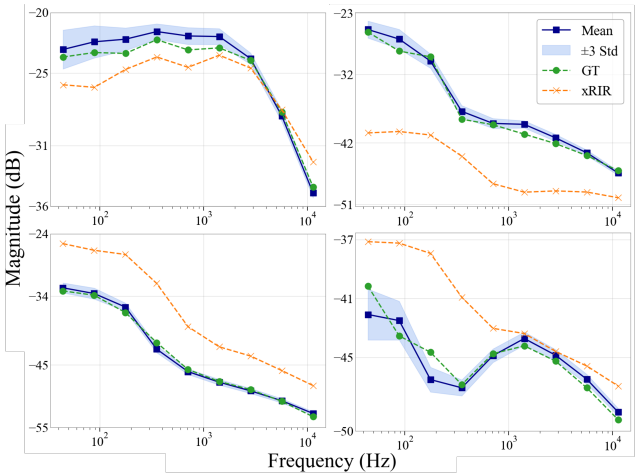


Figure 7. **Octave-band analysis of 4 RIRs in unseen rooms:** 100 samples per instance are generated with FLAC. The mean and ±3 standard deviation (covering 99.7% of the distribution) are shown. Std increases at low frequencies.

As shown in Tab. 1, FLAC maintains state-of-the-art performance in the one-shot setting, surpassing prior methods using eight recordings. Fig. 6 shows that FLAC remains significantly more stable than KNN and xRIR as $K$ decreases. Recall metrics are little affected by reduced acoustic observations, indicating that geometry provides the dominant cue for geometry-consistent RIR synthesis.

**Capturing uncertainty of few-shot RIR synthesis.** We study variability by generating 100 RIRs per conditioning, each produced with a different noise input. As shown by the octave-band analysis in Fig. 7, samples standard deviation increases at low frequencies. These bands also exhibit longer uncertainty persistence time, defined as the time until band-wise sample variance drops below the 75th percentile (see Fig. 8). This matches room acoustics theory: low-frequency responses are governed by sparse, boundary-dependent modes that are weakly constrained by limited context, whereas above the Schroeder frequency dense mode yields stable responses constrained by local geometry. This indicates that FLAC captures the inherent uncertainty of underconstrained few-shot settings. A deterministic variant (fixed noise) degrades performance (+6% T60, +10% C50, -40% R@5), confirming that stochasticity is essential. Quantitatively, FLAC's intra-conditioning diversity is $1.03_{\pm 0.20}$ vs. 22.96 between conditionings (a 4.5% ratio), showing that FLAC introduces meaningful stochasticity while remaining consistent with the context. See Appendix Ffor a t-SNE visualization.

Table 3. **Sim-to-real transfer to the Hearing-Anything-Anywhere dataset:** Few-shot methods are compared against Diff-RIR and INRAS, which require per-scene training (†). For FLAC, we report mean and standard deviation over 5 generations. With $K$=8, FLAC matches or exceeds xRIR and Diff-RIR on perceptual metrics, and with one-shot, it outperforms KNN and xRIR.

| Method | K | T60 (%) ↓ | C50 (dB) ↓ | EDT (ms) ↓ | R@5 (%) ↑ | $FD_G$ ↓ |
|---|---|---|---|---|---|---|
| Random Across Rooms | ✗ | 17.40 | 10.283 | 533.99 | 1.49 | 0.460 |
| Random Same Room | ✗ | 8.00 | 4.805 | 180.15 | 1.86 | 0.169 |
| Nearest Neighbor | 1 | 8.19 | 5.000 | 187.55 | 1.20 | **0.177** |
| xRIR | 1 | 8.63 | 4.862 | 183.27 | 14.85 | 0.363 |
| **FLAC** | 1 | **3.45**$_{\pm 0.02}$ | **2.170**$_{\pm 0.014}$ | **90.02**$_{\pm 0.24}$ | **17.94**$_{\pm 0.62}$ | 0.564 |
| Linear Interpolation | 8 | 4.12 | 2.695 | 88.19 | 3.62 | 0.904 |
| Nearest Neighbor | 8 | **2.89** | **1.923** | **77.24** | 9.61 | **0.169** |
| xRIR | 8 | 6.53 | 3.492 | 149.69 | **20.65** | 0.318 |
| **FLAC** | 8 | 3.10$_{\pm 0.01}$ | 2.167$_{\pm 0.004}$ | 84.52$_{\pm 0.24}$ | 17.41$_{\pm 0.59}$ | 0.585 |
| INRAS† | 12 | 6.61 | 3.966 | 158.07 | 2.27 | 0.797 |
| Diff-RIR† | 12 | 3.74 | 2.067 | 88.09 | 26.97 | 0.263 |

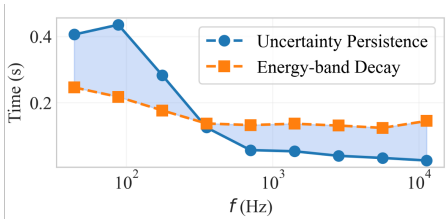


Figure 8. **Uncertainty persistence time** and band-wise energy decay, averaged over 100 unseen samples. Uncertainty lasts longer at low frequencies and decays faster at high frequencies.

**Sim-to-real transfer.** We evaluate real-world generalization on the HAA dataset [77]. Baselines include Diff-RIR [77], a physics-based differentiable renderer, and INRAS [74], a novel-view acoustic synthesis method, both trained with 12 references per room to predict RIRs at new locations. Unlike few-shot models, they must be retrained separately for each room, requiring hours of training. Following [49], we fine-tune xRIR and FLAC. Note that we do not fine-tune FLAC's VAE. Few-shot models adapt to all four rooms within minutes. For evaluation, AGREE is also fine-tuned on HAA. As shown in Tab. 3, with eight shots, FLAC outperforms xRIR and surpasses Diff-RIR on most perceptual metrics, despite using fewer references and no room-specific training. While 8-NN performs strongly, it copies existing RIRs, lacking spatial continuity (audible "jumps"). Remaining discrepancies likely stem from: (i) HAA's simplified geometry annotations (*e.g.*, tables as single planes); and (ii) the VAE not being fine-tuned on real recordings, which may cause its latent representation to miss certain acoustic phenomena. The small size of HAA proved insufficient for stable adaptation of the VAE.

Table 4. **Impact of geometry and acoustic encoders:** Performance on unseen AcousticRooms scenes using different configurations of the geometry $\phi_G$ and acoustic $\phi_A$ encoders. We compare xRIR's ViT and DINOv3 ViT-S/16 with three initialization strategies: trained from scratch, frozen, or fine-tuned ($\mathcal{W}_{\text{DINO}}$). For $\phi_A$, we evaluate the ResNet-18 and our frozen VAE encoder.

| $\phi_G$ ViT | $\mathcal{W}_{\text{DINO}}$ | $\phi_A$ | K | T60 (%) ↓ | C50 (dB) ↓ | EDT (ms) ↓ | R@5 (%) ↑ | $\text{FD}_G$ ↓ |
|---|---|---|---|---|---|---|---|---|
| [49] | ✗ | ResNet | 1 | 10.91 | 1.166 | 42.41 | 11.17 | 0.328 |
| S/16 [68] | ✗ | ResNet | 1 | 10.81 | 1.090 | 42.11 | 15.40 | 0.318 |
| S/16 [68] | ❄ | ResNet | 1 | 10.42 | 1.427 | 51.79 | 5.29 | 0.373 |
| S/16 [68] | ✓ | ResNet | 1 | 9.95 | **1.046** | 40.04 | **18.9** | **0.303** |
| S/16 [68] | ✓ | VAE | 1 | **9.40** | 1.057 | **39.31** | 17.11 | 0.310 |
| [49] | ✗ | ResNet | 8 | 9.46 | 1.063 | 39.57 | 11.80 | 0.333 |
| S/16 [68] | ✗ | ResNet | 8 | 9.29 | 0.994 | 38.61 | 16.24 | 0.320 |
| S/16 [68] | ❄ | ResNet | 8 | 8.87 | 1.298 | 46.41 | 5.92 | 0.378 |
| S/16 [68] | ✓ | ResNet | 8 | 8.60 | 0.970 | 37.13 | **19.38** | **0.305** |
| S/16 [68] | ✓ | VAE | 8 | **8.51** | **0.945** | **34.70** | 17.56 | 0.310 |

Table 5. **Impact of DiT variants:** Performance on unseen AcousticRooms scenes with In-Context, Cross-Attention (CA), and hybrid AdaLN+CA conditioning.

| Method | K | T60 (%) ↓ | C50 (dB) ↓ | EDT (ms) ↓ | R@5 (%) ↑ | $\text{FD}_G$ ↓ |
|---|---|---|---|---|---|---|
| In-Context | 1 | 69.68 | 11.199 | 1236.98 | 0.06 | 1.270 |
| CA | 1 | 15.68 | 1.750 | 85.98 | 6.10 | 0.424 |
| AdaLN+CA | 1 | **9.95** | **1.046** | **40.04** | **18.92** | **0.303** |
| In-Context | 8 | **8.12** | 1.081 | 41.97 | 0.194 | 0.316 |
| CA | 8 | 9.31 | 1.234 | 45.81 | 11.93 | 0.342 |
| AdaLN+CA | 8 | 8.60 | **0.970** | **37.13** | **19.38** | **0.305** |

Yet, FLAC one-shot outperforms both KNN and eight-shot xRIR, highlighting its advantage in data-scarce conditions.

**Perceptual Evaluation.** We conducted a listening study with 46 participants on 14 unseen AR scenes. Participants were presented with the ground-truth, audio generated by FLAC (1-shot) and xRIR (8-shot), and were asked to select which audio sounded closer to the GT. FLAC was preferred in 93.01% of cases. Details are given in Appendix G.

### 5.6. Ablation Study

**Conditioning modalities.** We analyze the impact of each conditioning modality by removing either geometry or audio (see Tab. 1). When conditioned only on geometry, the model maintains strong audio-to-audio recall and outperforms random RIR prediction, confirming that geometric cues provide rich information for RIR synthesis. In contrast, using only audio leads to a drop in geometry-related metrics (recall and $\text{FD}_G$). For perceptual metrics, geometry-only achieves higher C50 and EDT but lower T60 compared to the audio-only version. This aligns with their physical meaning: C50 and EDT are influenced by early reflections from nearby surfaces, while T60 captures global reverberation that is harder to infer from local geometry. Overall, combining both modalities through cross-attention yields the best results, demonstrating the complementary nature of geometric and acoustic conditioning.

**Geometry conditioning encoder.** In Tab. 4 we study how the choice of geometry conditioning encoder affects performance. We compare the ViT architecture from xRIR with DINOv3 ViT-S/16, which have similar parameter counts (19.8M vs. 21.7M). For DINOv3, we test three variants: (i) trained from scratch, (ii) frozen pretrained weights, and (iii) fine-tuned jointly with the model. Even when trained from scratch, the ViT-S/16 outperforms xRIR's ViT. As our input differs substantially from RGB images, freezing DINO weights degrades performance. Fine-tuning DINO yields the best overall results. We report a similar analysis for the AGREE geometric encoder in Appendix C.3 , where fine-tuning DINOv3 ViT-S/16 consistently improves zero-shot retrieval. Note that even with the same conditioning architecture as xRIR, one-shot FLAC achieves comparable T60, $\text{FD}_G$ and higher C50, EDT and R@5 than eight-shot xRIR.

**Acoustic conditioning encoder.** We evaluate replacing the jointly trained ResNet-18 with our frozen, pretrained VAE encoder (see Tab. 4). The VAE improves cross-room generalization, though at higher computational cost. For efficiency, we use the ResNet-18 as the default encoder.

**DiT variants.** We study different DiT conditioning strategies (see Tab. 5). *In-Context* concatenates all conditioning information with the input before self-attention. *Cross-Attention* applies conditioning solely via cross-attention layers. Our approach (see Fig. 3) injects target information through AdaLN, and contextual information via cross-attention. *AdaLN+CA* outperforms alternative designs. Illustrations of variants are provided in Appendix B.3.

## 6. Conclusion

We introduced FLAC, a generative approach for few-shot acoustic synthesis based on flow matching. By conditioning generation on multimodal few-shot context, FLAC can synthesize RIRs at arbitrary sensor positions in novel environments. Our method captures the inherent ambiguity of few-shot RIR synthesis, an aspect overlooked by existing deterministic methods. Experiments on two datasets demonstrated state-of-the-art performance in novel environments, even with a single reference RIR. We also introduced AGREE, a joint-embedding space between RIRs and geometry enabling both zero-shot cross-modal retrieval and geometry-consistency evaluation. FLAC produces RIRs that are both perceptually accurate and consistent with the scene, an important aspect for immersive virtual experiences. Future work may include supporting multiple sample rates in a single model, and collecting a larger, more diverse real-world audio-visual dataset to improve sim-to-real transfer. The AGREE embedding could also benefit broader audio-visual learning tasks.

## Acknowledgments

This work was supported by the French Agence Nationale de la Recherche (ANR), under grant ANR22-CE94-0003 and was granted access to the HPC resources of IDRIS under the allocation 2024-AD011015475R1 made by GENCI. We would like to thank Simon de Moreau and the anonymous reviewers for their insightful comments and suggestions.

## References

[1] Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023. 3

[2] 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 *NeurIPS*, 2024. 1, 2

[3] Amandine Brunetto, Sascha Hornauer, X Yu Stella, and Fabien Moutarde. The audio-visual batvision dataset for research on sight and sound. In *IROS*, 2023. 2

[4] Amandine Brunetto, Sascha Hornauer, and Fabien Moutarde. NeRAF: 3D scene infused neural radiance and acoustic fields. In *ICLR*, 2025. 1, 2, 5

[5] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigaton in 3d environments. In *ECCV*, 2020. 5

[6] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *ICLR*, 2021. 2

[7] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *CVPR*, 2022. 2

[8] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS*, 2022. 5

[9] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. Learning audio-visual dereverberation. In *ICASSP*, 2023. 2

[10] Mingfei Chen and Eli Shlizerman. Av-cloud: Spatial audio rendering through audio-visual cloud splatting. In *NeurIPS*, 2024. 1, 2

[11] Ziyang Chen, Shengyi Qian, and Andrew Owens. Sound localization from motion: Jointly learning sound direction and camera rotation. In *ICCV*, 2023. 2

[12] Ziyang Chen, Israel D. Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. In *CVPR*, 2024. 2

[13] Sanjoy Chowdhury, Sreyan Ghosh, Subhrajyoti Dasgupta, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. In *ICCV*, 2023. 2

[14] Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see 3d spatial layout with two ears. In *ICRA*, 2020. 2

[15] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022. 4

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 4

[17] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP*, 2023. 2

[18] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2

[19] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Long-form music generation with latent diffusion. *arXiv preprint arXiv:2404.10301*, 2024. 2, 4

[20] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024. 2, 4

[21] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020. 2

[22] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *CVPR*, 2019. 2

[23] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *ECCV*, 2020. 2

[24] Ruohan Gao, Hao Li, Gokul Dharan, Zhuzhu Wang, Chengshu Li, Fei Xia, Silvio Savarese, Li Fei-Fei, and Jiajun Wu. Sonicverse: A multisensory simulation platform for embodied household agents that see and hear. In *ICRA*, 2023. 2

[25] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Visually-guided audio spatialization in video with geometry-aware multi-task learning. *IJCV*, 2023. 2

[26] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023. 2

[27] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 2

[28] Yiwei Guo, Chenpeng Du, Ziyang Ma, Xie Chen, and Kai Yu. Voiceflow: Efficient text-to-speech with rectified flow matching. In *ICASSP*, 2024. 2

[29] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP*, 2022. 2

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 5

[32] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPSW*, 2021. 3

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[34] Diwei Huang, Kunyang Lin, Peihao Chen, and Qing Du. Map-guided few-shot audio-visual acoustics modeling. In *ICASSP*, 2025. 1, 2

[35] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*, 2023. 2

[36] Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*, 2024. 2, 4

[37] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):139–1, 2023. 2

[38] Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. Visage: Video-to-spatial audio generation. In *ICLR*, 2025. 2

[39] Sang-gil Lee, Zhifeng Kong, Arushi Goel, Sungwon Kim, Rafael Valle, and Bryan Catanzaro. ETTA: Elucidating the design space of text-to-audio models. In *ICML*, 2025. 2

[40] Mark Levy, Bruno Di Giorgi, Floris Weers, Angelos Katharopoulos, and Tom Nickson. Controllable music production with diffusion models and guidance gradients. *arXiv preprint arXiv:2311.00613*, 2023. 4

[41] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. In *NeurIPS*, 2023. 1, 2, 5

[42] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *ICCVW*, 2023. 2

[43] Susan Liang, Dejan Markovic, Israel D Gebru, Steven Krenn, Todd Keebler, Jacob Sandakly, Frank Yu, Samuel Hassel, Chenliang Xu, and Alexander Richard. Binauralflow: A causal and streamable approach for high-quality binaural speech synthesis with flow matching models. In *ICML*, 2025. 2

[44] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 2, 3

[45] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *ICML*, 2023. 2

[46] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 2

[47] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022. 3

[48] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 3

[49] Xiulong Liu, Anurag Kumar, Paul Calamia, Sebastià V. Amengual Garí, Calvin Murdock, Ishwarya Ananthabhotla, Philip Robinson, Eli Shlizerman, Vamsi Krishna Ithapu, and Ruohan Gao. Hearing anywhere in any environment. In *CVPR*, 2025. 1, 2, 4, 5, 6, 7, 8

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[51] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. In *NeurIPS*, 2022. 1, 2

[52] Tanvir Mahmud, Shentong Mo, Yapeng Tian, and Diana Marculescu. Ma-avt: Modality alignment for parameter-efficient audio-visual transformers. In *CVPR*, 2024. 2

[53] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. *arXiv preprint arXiv:2404.09956*, 2024. 2

[54] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. *NeurIPS*, 2022. 1, 2, 4, 5

[55] Sagnik Majumder, Hao Jiang, Pierre Moulon, Ethan Henderson, Paul Calamia, Kristen Grauman, and Vamsi Krishna Ithapu. Chat2map: Efficient scene mapping from multi-ego conversations. In *CVPR*, 2023. 2

[56] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *CACM*, 2021. 2

[57] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *NeurIPS*, 31, 2018. 2

[58] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *CVPR*, 2021. 2

[59] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond image to depth: Improving depth prediction using echoes. In *CVPR*, 2021. 2

[60] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 4

[61] Senthil Purushwalkam, Sebastia Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. In *ICCV*, 2021. 2

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 5

[63] Anton Ratnarajah, Shi-Xiong Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. Fast-rir: Fast neural diffuse room impulse response generator. In *ICASSP*, 2022. 6

[64] Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. Av-rir: Audio-visual room impulse response estimation. In *CVPR*, 2024. 2

[65] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015. 2

[67] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *NeurIPS*, 2016. 4

[68] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 4, 8

[69] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *ICCV*, 2021. 2

[70] Arjun Somayazulu, Changan Chen, and Kristen Grauman. Self-supervised visual acoustic matching. *NeurIPS*, 2024. 2

[71] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[72] Christian J Steinmetz and Joshua D Reiss. auraloss: Audio focused loss functions in pytorch. In *Digital music research network one-day workshop*, 2020. 4

[73] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 4

[74] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. In *NeurIPS*, 2022. 1, 2, 5, 7

[75] Zhenyu Tang, Rohith Aralikatti, Anton Jeran Ratnarajah, and Dinesh Manocha. Gwa: A large high-quality acoustic dataset for audio processing. In *SIGGRAPH*, 2022. 5

[76] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic object prediction and spatial sound super-resolution with binaural sounds. In *ECCV*, 2020. 2

[77] Mason Long Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *CVPR*, 2024. 1, 2, 5, 7

[78] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP*, 2020. 4

[79] Abdelrahman Younes, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds. *RA-L*, 2023. 2

[80] Wenjie Zhang, Jun Yin, Long Ma, Peng Yu, Xiaoheng Jiang, Zhen Tian, and Mingliang Xu. Echodiffusion: Waveform conditioned diffusion models for echo-based depth estimation. In *AAAI*, 2025. 2

[81] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *ECCV*, 2020. 2

[82] Lingyu Zhu, Esa Rahtu, and Hang Zhao. Beyond visual field of view: Perceiving 3d environment with echoes and vision. *CVPRW*, 2022. 2

[83] Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. In *NeurIPS*, 2020. 4