π-AVAS: Can Physics-Integrated Audio-Visual Modeling Boost Neural Acoustic Synthesis?

Susan Liang, Chao Huang, Yunlong Tang, Zeliang Zhang, Chenliang Xu
University of Rochester

Abstract

The Audio-Visual Acoustic Synthesis (AVAS) task aims to model realistic audio propagation behavior within a specific visual scene. Prior works often rely on sparse image representations to guide acoustic synthesis. However, we argue that this approach is insufficient to capture the intricate physical properties of the environment and may struggle with generalization across diverse scenes. In this work, we review the limitations of existing pipelines and address the research question: Can we leverage physical audio-visual associations to enhance neural acoustic synthesis? We introduce Physics-Integrated Audio-Visual Acoustic Synthesis (PI-AVAS or π-AVAS), a novel framework designed with two key objectives. i) Generalization: We develop a vision-guided audio simulation framework that leverages physics-based sound propagation. By explicitly modeling vision-grounded geometry and sound rays, our approach achieves robust performance across diverse visual environments. ii) Realism: While simulation-based approaches offer generalizability, they often compromise on realism. To mitigate this, we incorporate a second stage for data-centric refinement, where we propose a flow matching-based audio refinement model to narrow the gap between simulation and real-world audio-visual scenes. Extensive experiments demonstrate the effectiveness and robustness of our method. We achieve state-of-the-art performance on the RWAVS-Gen, RWAVS, and RAF datasets. Additionally, we show that our approach can be seamlessly integrated with existing methods to significantly improve their performance.

1. Introduction

The audio-visual acoustic synthesis task, as introduced by Chen et al. [6] and Liang et al. [21], aims to generate realistic binaural audio for new speaking and listening positions based on vision data. This task presents unique challenges, including synthesizing realistic binaural audio and achieving novel-view synthesis. Various approaches have been proposed to address this problem [6, 9, 21, 22, 26, 43]. NAF [26] simulates sound propagation within a scene using a local feature grid and an implicit decoder. INRAS [43] divides sound modeling into a three-stage implicit neural field. Chen et al. [6] present a vision-conditioned audio transformation network to synthesize audio for new listening positions. Liang et al. [21] design a NeRF-like system to jointly render audio and visual content.

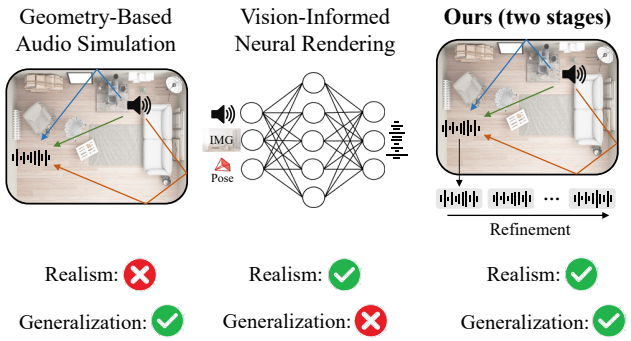


Figure 1. An intuitive comparison between audio simulation approaches, neural rendering approaches, and our π-AVAS. Physics-based approaches generalize well but lack realism, while neural rendering produces high-quality results but struggles with novel sound sources. Our two-stage method achieves both realism and generalization using the physics-based sound simulation and audio refinement model.

While these methods produce plausible results for fixed sound sources, they struggle to generalize to novel sound sources due to inadequate modeling of sound propagation. Sound propagation is influenced by various environmental factors, including the positions of the sound emitter and receiver, the room geometry, and the material properties of surfaces. The complex interactions among these physical properties impact the propagation behavior in a given environment. To address these factors, existing methods typically learn from sparse images implicitly within neural networks. Although this implicit modeling of visual information can support audio synthesis, it falls short of accurately capturing the necessary physical properties that govern sound propagation. In other words, the limited image-audio pairs do not provide sufficient information to infer the

complete room geometry and scene layout. Consequently, these models experience significant performance degradation when confronted with novel sound sources.

Recognizing the limitations of implicit learning approaches in capturing physical audio-visual associations, we investigate the question: *Can we leverage explicit physical priors in audio-visual modeling to bridge this gap?* In this paper, we propose a novel two-stage method to address this challenge. In the first stage, we design a vision-guided sound simulation framework that improves generalization for novel emitter and receiver positions. We begin by reconstructing a 3D mesh environment from a set of images using NeRF [29] or Gaussian Splatting [18]. This scene mesh effectively captures the geometry and structure of the environment, which are critical factors influencing sound propagation. For a given sound source and receiver pair, we model sound propagation within the mesh scene by treating sound as a ray [36]. This physics-based simulation approach enables robust generalization to new positions, even in complex environments.

While explicit modeling of scenes from visual input provides a significant advantage in generalization, it falls short of achieving realistic audio rendering [4, 46] for several reasons. (1) The material properties of each bounce surface, such as absorption, scattering, and transmission, are not fully considered. (2) The simulation traces only a limited number of sound rays due to computational restrictions, and low-frequency values are imprecise. (3) Background noise effects are not modeled. To address these issues, we propose our second stage to enhance the realism of the simulated sound. We utilize a conditional flow matching model [25, 47] to refine the coarsely simulated sound. With the powerful generative capability of flow-matching models, we can effectively correct errors from the first stage. An intuitive comparison between traditional audio simulation methods, neural rendering approaches, and our proposed method is presented in Fig. 1.

We conduct extensive experiments on three real-world datasets: RWAVS-Gen, RWAVS [21], and RAF [9] datasets. RWAVS-Gen and RWAVS datasets measure the waveform sound generation quality, and the RAF dataset assesses the room impulse response rendering performance. The experimental results demonstrate that our method exhibits strong generalization and can be readily applied to novel sound sources. We also show that our method can be integrated into existing approaches to improve their generalization performance. In conclusion, our contributions are as follows:

- We introduce a novel physics-integrated audio-visual acoustic synthesis framework to generate realistic audio content at novel positions based on visual information.
- We propose a vision-guided audio simulation method to enhance generalization for novel sources and listeners.
- We design a flow matching-based audio refinement model to bridge the gap between simulation sounds and real-world recordings.
- Our experiments highlight the limitations of existing approaches, demonstrate the advantages of our method, and show the applicability of our model.

2. Related Work

Our work is closely related to areas such as vision-informed audio generation and flow matching models. We discuss each area and related work in the following section.

2.1. Vision-Informed Audio Generation

Vision-informed audio generation focuses on synthesizing audio based on visual inputs like images, videos, meshes, and poses. Many studies propose neural network-based audio generation pipelines [5, 6, 12, 15, 21, 23, 26, 27, 32, 33, 43, 52]. For instance, 2.5D Visual Sound [12] employs a U-Net [34] to synthesize binaural audio conditioned on an image, while Chen et al. [5] introduce a cross-modal transformer [49] to generate audio that matches room acoustics informed by an image. Some researchers have also developed video-guided audio generation methods, such as Diff-foley [27], and Movie Gen [32], which produce synchronized audio by considering temporal cues in videos.

Another key area of vision-informed audio generation is pose-conditioned audio rendering, which is the focus of this paper. Inspired by Neural Radiance Field (NeRF) [29], several works [1, 2, 6, 7, 11, 14, 20, 21, 26, 33, 43] explore novel-pose audio synthesis by learning an audio field. Chen et al. [6] introduce a CNN-based network for transforming audio to synthesize sound at novel poses, while Liang et al. [21] design a NeRF-like system to jointly generate audio and visual content conditioned on poses. Although these pose-conditioned approaches generate plausible results for fixed sound sources, they face challenges with novel sound sources. In comparison, our method can easily render sound for new sources, thanks to our vision-guided audio simulation approach with physics integration.

2.2. Flow Matching Models

Deriving from Continuous Normalizing Flows (CNFs) [8], Lipman et al. [25] introduce Flow Matching to train CNFs in a simulation-free manner. Flow Matching, especially Optimal Transport Flow Matching [28], models the transformation between noise and data samples in a simpler and more efficient way than diffusion models [13, 41], leading to more stable training and better performance. Tong et al. [47] extend Flow Matching to arbitrary distribution transformations, including probability paths between different data distributions [24]. Inspired by this, we treat our simulated audio as one distribution and the target binaural audio as another distribution and use Flow Matching to refine

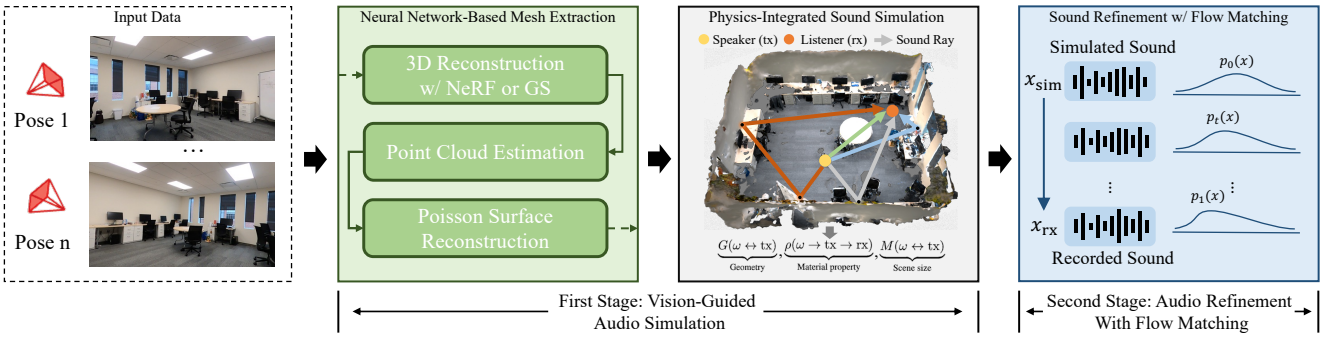


Figure 2. Overview of our approach. Our framework π-AVAS consists of two stages: the vision-guided audio simulation and the audio refinement with flow matching. In the first stage, we conduct 3D scene reconstruction and simulate sound propagation between a speaker and a listener in the mesh scene. In the second stage, we refine the coarsely simulated sound with a flow matching model, enhancing the quality of the synthesized sound.

the simulation results. The audio refinement network effectively corrects simulation errors from the first stage.

3. Method

Our method aims to boost the performance of audio-visual acoustic synthesis models for novel sound sources. We design a novel Physics-Integrated Audio-Visual Acoustic Synthesis model (PI-AVAS or π-AVAS) with two stages. In the first stage (Sec. 3.1), we introduce a vision-guided audio simulation approach that integrates the physical properties of sound propagation. In the second stage (Sec. 3.2), we propose a conditional flow matching model that refines the prediction results from the first stage, improving the quality of audio generation. Additionally, we introduce a data augmentation strategy in Sec. 3.3 to facilitate model training. The complete pipeline is illustrated in Fig. 2.

3.1. Vision-Guided Audio Simulation

To enhance the generalization for audio sources and receivers in novel positions, we design a vision-guided audio simulation framework.

3D Scene Mesh Extraction. We aim to reconstruct meshes of an audio-visual scene from input images in the first step (see the second subfigure in Fig. 2). Given a set of images and their corresponding camera poses, we utilize Neural Radiance Field (NeRF) [29, 45] or 3D Gaussian Splatting [18] to learn a neural representation of the given environment. Then, we convert the NeRF model weights or Gaussian points to 3D point clouds, which are a more compatible data format. Once the point clouds are reconstructed, we use Poisson surface reconstruction [17] to generate meshes of the audio-visual scene. Empirically, we do not observe a noticeable difference between the reconstructed 3D meshes of NeRF and Gaussian Splattings. We leave a detailed comparison between NeRF, Gaussian Splatting, and traditional Structure-from-Motion approaches [38] for future work, as this is not the main focus of our paper.

Physics-Integrated Sound Simulation. After we generate the meshes of an audio-visual scene, we can simulate the audio propagation between arbitrary sound emitters and sound receivers in this scene (see the third subfigure in Fig. 2).

Specifically, for a pair of sound source tx and sound receiver rx, we treat the sound emitted by the source $x_{\mathrm{tx}} \in \mathbb{R}^n$ (n is the audio length) as a collection of rays, tracing each ray's interaction with the room's meshes [3, 40]. We denote the energy received at the receiver from the transmitter as $E(\mathrm{tx} \to \mathrm{rx})$, from which we omit the time delay variable for simplicity. We model both direct propagation and indirect reflection to calculate the received energy:

$$E(\mathrm{tx} \to \mathrm{rx}) = \underbrace{E_{\mathrm{d}}(\mathrm{tx} \to \mathrm{rx})}_{\text{Direct propagation}} + \underbrace{E_{\mathrm{id}}(\mathrm{tx} \to \mathrm{rx})}_{\text{Indirect reflection}}. \quad (1)$$

$E_{\mathrm{d}}(\mathrm{tx} \to \mathrm{rx})$ is the energy of direct propagation and $E_{\mathrm{id}}(\mathrm{tx} \to \mathrm{rx})$ is the reflected energy defined as

$$E_{\mathrm{id}}(\mathrm{tx} \to \mathrm{rx}) = \int_{\Omega} \underbrace{E(\omega \to \mathrm{tx})}_{\text{Energy}} \underbrace{G(\omega \leftrightarrow \mathrm{tx})}_{\text{Geometry}} \underbrace{\rho(\omega \to \mathrm{tx} \to \mathrm{rx})}_{\text{Material property}} * \underbrace{M(\omega \leftrightarrow \mathrm{tx})}_{\text{Scene size}} d\omega, \quad (2)$$

where Ω is the entire mesh space, ω is an area of Ω, $M(\omega \leftrightarrow \mathrm{tx})$ measures energy absorption and time delay, $G(\omega \leftrightarrow \mathrm{tx})$ means the energy dispersion and occlusion during sound propagation, $\rho(\omega \to \mathrm{tx} \to \mathrm{rx})$ represents the acoustic property of each bounce surface, and the asterisk $*$ is the convolution operation.

Eq. (1) can be converted to an infinite sum of integrals and solved with the Neumann series expansion [19]. We then calculate the impulse response based on the accumulated energy E and convolve that with the sound source to generate the simulated sound. To improve the realism of the

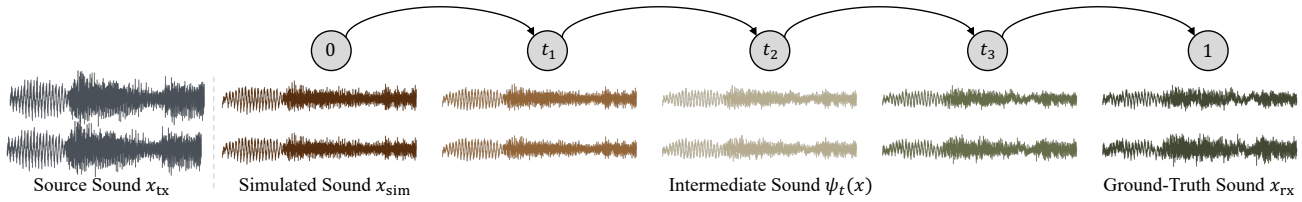


Figure 3. Visualization of the audio refinement process with the flow matching model. The flow matching model learns a vector field that gradually transforms a simulated sound x_{sim} to a ground-truth sound x_{rx}.

simulated sound, we also apply the Head-Related Transfer Function (HRTF) to generate binaural audio $x_{\text{sim}} \in \mathbb{R}^{2\times n}$ based on the listener's head direction. Because it is challenging to estimate material property solely based on visual information [10, 39], we assign default material coefficients to all surfaces.

So far, we simulate sound propagation between the sound source and the receiver using visual information obtained from a set of images. Since this approach integrates the physical properties of sound propagation, it generalizes well to novel sound sources and listeners.

3.2. Audio Refinement With Flow Matching

Although the vision-guided audio simulation approach offers robustness for novel poses, there is a noticeable fidelity gap between simulated and recorded sounds [4, 46] as mentioned in the introduction section, e.g., the material property is not correctly modeled. Therefore, we introduce our second stage to further enhance the quality of the generated sound x_{sim}. We propose a flow matching model to refine and improve the simulated audio. We illustrate the audio refinement process with the flow matching model in Fig. 3.

Flow Matching Formula. Specifically, we treat the simulated sound $x_{\text{sim}} \in \mathbb{R}^{2\times n}$ as one data distribution $\mathcal{N}(x|x_{\text{sim}}, \sigma^2 I)$ and the recorded sound (ground-truth sound) $x_{\text{rx}} \in \mathbb{R}^{2\times n}$ as another data distribution $\mathcal{N}(x|x_{\text{rx}}, \sigma^2 I)$, where σ is a predefined standard deviation. We model the sound refinement process as a transfer between these two distributions. We design the following time-dependent probability path $p_t : [0,1] \times \mathbb{R}^n \to \mathbb{R}_{>0}$:

$$p_t(x) = \mathcal{N}(x|tx_{\text{rx}} + (1-t)x_{\text{sim}}, \sigma^2 I), \tag{3}$$

where $p_0(x) = \mathcal{N}(x|x_{\text{sim}}, \sigma^2 I)$, $p_1(x) = \mathcal{N}(x|x_{\text{rx}}, \sigma^2 I)$, and time step $t \in [0,1]$. We define the time-dependent flow $\psi_t(x) : [0,1] \times \mathbb{R}^n \to \mathbb{R}^n$ as follows:

$$\psi_t(x) = tx_{\text{rx}} + (1-t)x_{\text{sim}} + \sigma\epsilon, \tag{4}$$

where ϵ is sampled from a standard Gaussian distribution $\mathcal{N}(0, I)$. According to the definition of vector fields, we can derive the vector field $v_t : [0,1] \times \mathbb{R}^n \to \mathbb{R}^n$ using the flow defined in Eq. (4):

$$v_t(\psi_t(x)) = \frac{d}{dt}\psi_t(x) = x_{\text{rx}} - x_{\text{sim}}. \tag{5}$$

Then we train a deep neural network $u_t(\psi_t(x), x_{\text{tx}}, p; \theta)$ to fit the vector field defined in Eq. (5), where θ is the trainable parameters of a neural network, x_{tx} is the source sound, and p is the pose information of sound source tx and listener rx. The training objective is

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{q_0(x),q_1(x),t} \left\| u_t(\psi_t(x), x_{\text{tx}}, p; \theta) - (x_{\text{rx}} - x_{\text{sim}}) \right\|^2. \tag{6}$$

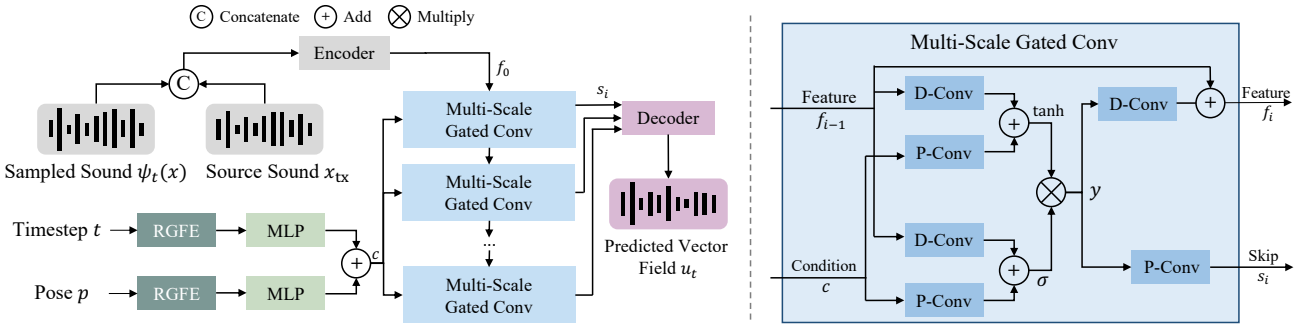
We provide pseudo-code for training the audio refinement model in the appendix. In our experiments, we use Short-time Fourier transform (STFT) to transform both the target $x_{\text{rx}} - x_{\text{sim}}$ and predicted $u_t(\psi_t(x), x_{\text{tx}}, p; \theta)$ vector fields from a time space to a time-frequency space and calculate the L2 distance as the training loss.

After training, we utilize the network for audio refinement. Given a simulated sound x_{sim}, we generate an enhanced sound x_{rx} by solving the ordinary differential equation:

$$\begin{aligned} \frac{d}{dt}\psi_t(x) &= u_t(\psi_t(x), x_{\text{tx}}, p; \theta), \\ \psi_0(x) &= x_{\text{sim}}. \end{aligned} \tag{7}$$

The solution $\psi_1(x)$ is the synthesized binaural audio. In this paper, we study one first-order solver (Euler solver) and two second-order solvers (Midpoint solver and Heun solver). We compare different solvers and provide the pseudo-code of inference in the appendix.

Audio Refinement Network. In Fig. 4 (a), we show the architecture of our network that is designed to approximate the vector field $v_t(x)$. Given an intermediate sound $\psi_t(x)$ sampled from Eq. (4), we concatenate it with the source sound x_{tx} emitted by the loudspeaker. The concatenated sound is first fed to a two-layer convolutional encoder to enrich the channel dimension and then passed to a stack of multi-scale gated convolution blocks. To condition the vector field prediction on the source and listener's poses p and time step t, we first project them into a high-frequency space using Random Gaussian Fourier Embedding (RGFE) [44], followed by MLPs. The resulting embeddings form the flow matching conditions c, which are then passed to the multi-scale gated convolution blocks. Each multi-scale gated convolution block uses condition c to adjust the sound features f_{i-1} and predicts the next sound features f_i and a skip feature s_i. Finally, we combine all



(a) Audio Refinement Network

(b) Multi-Scale Gated Convolution Block

Figure 4. Model architecture. (a) shows the audio refinement network. Given an intermediate sound $\psi_t(x)$ and a source sound x_{tx}, we concatenate them and pass them through an encoder followed by a series of multi-scale gated convolution blocks. We use timestep t and pose p as conditions by encoding them with RGFE and MLPs. All skip features s_i are gathered and used to predict the vector field u_t. (b) illustrates the multi-scale gated convolution block. We design a gate operator to filter and control the intermediate features f_i.

skip features $\{s_0, s_1, \ldots, s_{L-1}\}$ together, where L is the number of blocks, and use a two-layer convolutional decoder to predict the vector field u_t.

We present our multi-scale gated convolution block in Fig. 4 (b). Inspired by WaveNet [48] and ViGAS [6], we design a gate operator to filter and control intermediate features. Given a feature f_{i-1} from the previous block, we utilize dilated convolution layers (D-Conv) to learn meaningful features $D_1(f_{i-1})$ and $D_2(f_{i-1})$, where D_1 and D_2 are dilated convolution layers. We also use pointwise convolution layers (P-Conv) to extract condition features $P_1(c)$ and $P_2(c)$, where P_1 and P_2 are pointwise convolution layers. We apply the gated operator using the following equation:

$$y = \tanh(D_1(f_{i-1}) + P_1(c)) \otimes \sigma(D_2(f_{i-1}) + P_2(c)), \quad (8)$$

where $\otimes$ means Hadamard product.

Then we feed y through a pointwise convolution layer to form the residual and add it to the input feature f_{i-1} to generate the new feature f_i. We feed y through another pointwise convolution layer to generate a skip feature s_i. Since we gradually increase the dilation size of each block, we name it the multi-scale gated convolution block.

3.3. Data Augmentation

Data scarcity presents a challenge when we train audio-visual acoustic synthesis models, as each scene is typically recorded in a 10 to 20-minute video. This yields only 600 to 1200 samples (at 1 fps) per scene for training. To facilitate training, we propose a data augmentation strategy (see Fig. 5). Given that the video captures continuous camera movement, we approximate the camera's entire trajectory by interpolating between discrete training poses (shown as orange poses in the figure). We then shift each training pose randomly along this trajectory by up to one second forward or backward. These shifted positions serve as augmented poses (blue poses), and we pair them with temporally corresponding audio clips as augmented sound samples.

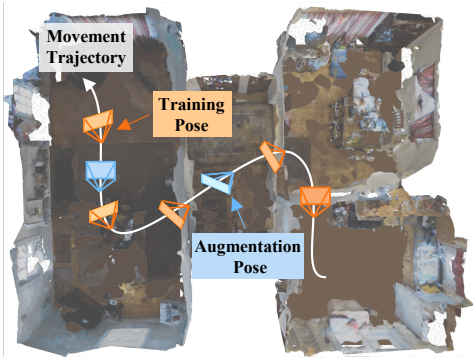


Figure 5. Data augmentation. We randomly shift each training pose, represented in orange, along the interpolated camera trajectory by up to one second forward or backward. The shifted poses, shown in blue, serve as augmented poses to aid in model training.

4. Experiments

4.1. Generalization Evaluation

We first evaluate the generalization ability of various methods to novel sound sources.

Experiment Setup. We use the Real-World Audio-Visual Scene (RWAVS) dataset [21] to benchmark our method. The RWAVS dataset captures multimodal data, including the source position, camera poses, mono source sounds, and binaural received sounds for each scene. The dataset captures data from diverse environments, such as offices, multi-room houses, apartments, and outdoor spaces. For each scene, the original RWAVS dataset contains multiple videos with varying sound source locations. To examine generalization issues in existing approaches, we design a

Table 1. Quantitative comparison with state-of-the-art methods on the RWAVS-Gen dataset using the generalization evaluation setup. We also show the inference speed and the model size of all methods. We highlight the best result in bold.

Methods	Office		House		Apartment		Outdoors		Overall		Speed (ms)	Size (MB)
	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓		
INRAS [43]	2.126	0.182	3.605	0.220	4.535	0.232	2.058	0.157	3.081	0.198	2.9	0.79
NAF [26]	2.275	0.181	2.873	0.186	4.878	0.231	1.575	0.135	2.900	0.183	4.6	0.74
ViGAS [6]	2.137	0.183	3.878	0.213	3.946	0.221	1.967	0.154	2.982	0.193	12.8	9.72
AV-NeRF [21]	2.086	0.180	3.759	0.221	4.520	0.230	2.308	0.165	3.168	0.199	2.1	3.04
π-AVAS (Ours)	**1.856**	**0.163**	**1.946**	**0.140**	**3.898**	**0.209**	**1.326**	**0.125**	**2.257**	**0.159**	10.4	5.62

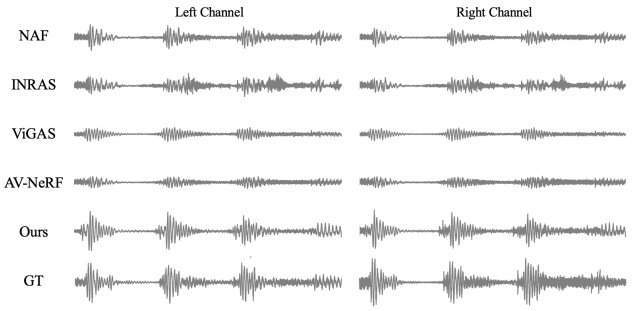


Figure 6. Visualization of synthesized binaural sounds for novel sound sources and listeners.

new evaluation setup: in each environment, we select one video as training data, with the remaining videos used for evaluation. This setup allows us to effectively assess how well existing methods generalize to new sound source locations within the same environment. We present an example in the appendix. We denote this new benchmark as RWAVS-Gen to distinguish it from the original RWAVS benchmark.

Following AV-NeRF [21], we select NAF [26], INRAS [43], ViGAS [6], and AV-NeRF as our baselines. NAF models sound propagation within a scene by using a local feature grid and an implicit decoder. INRAS disentangles sound modeling through a three-stage implicit neural field. ViGAS predicts sound at a new location by leveraging audio-visual features from source viewpoints. AV-NeRF constructs a multi-modal neural field to condition audio modeling on 3D visual scene context.

We choose magnitude distance (MAG) [51] and envelope distance (ENV) [30] metrics to evaluate audio quality following RWAVS [21].

Quantitative Results. We compare our model with existing approaches on the RWAVS-Gen dataset and report their generalization performance in Tab. 1. All methods show better audio rendering performance in the office and the outdoor scenes while causing worse audio generation in the house and apartment scenes because the office is a classic shoebox environment with four parallel walls and the outdoor scene is an open space without occlusion. Our model achieves the best performance across all four environments, achieving the lowest metric losses, with 2.257 MAG and 0.159 ENV. This demonstrates the strong robustness and generalization capability of our physics-integrated model to novel sound sources.

Inference Speed. We test the inference speed of different models to render one second of binaural audio with one RTX 4090 GPU and report the results in the rightmost column in Tab. 1. Since our approach involves audio simulation (2.1ms) and requires 4 steps to complete flow matching estimation (8.3ms), it is slower than some methods that directly output sound. However, (1) our approach still meets the real-time requirement (16.7ms for 60 FPS and 33.3ms for 30 FPS), meaning it causes no noticeable delay in real-world audio applications. (2) There are many successful works we can use to reduce inference time without performance degradation, such as consistency models [42] and adversarial diffusion distillation [35]. If we reduce the number of steps to 1, our method needs only 4.2 ms for inference.

Qualitative Comparison. We visualize the generated sounds of different approaches in Fig. 6 for an intuitive comparison. As depicted, existing methods encounter challenges when applied to novel sound sources, resulting in inaccurate sound volume. In contrast, our physics-based approach effectively considers the changes in sound source locations and accurately produces binaural audio.

4.2. Standard Evaluation

We then compare the rendering performance of our π-AVAS model with other methods using standard benchmarks.

Experiment Setup. We use the original RWAVS and Real Acoustic Field (RAF) datasets to benchmark our approach. For the RWAVS dataset, we make no modification to the benchmark and use the original setup to test our model. The RAF dataset is a real-world room impulse response dataset that densely captures real room acoustic data. It provides impulse response signals of an office environment in two conditions: empty and furnished. It allows to study the difference in acoustic fields introduced by furniture.

Besides the baselines used in the RWAVS-Gen experiment, we include state-of-the-art approaches on RWAVS and RAF datasets to augment our experiment. AV-GS [1] introduces an audio-visual Gaussian Splatting method that explicitly represents a scene for acoustic synthesis. SOAF [11] designs an occlusion-aware acoustic field. AVR [20] utilizes the volume rendering technique to generate acoustic impulse responses.

Table 2. Quantitative comparison with state-of-the-art methods on the original RWAVS dataset. We highlight the best-performing result in bold and underline the second-best result.

Methods	Office		House		Apartment		Outdoors		Overall	
	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓
Mono-Mono	9.269	0.411	11.889	0.424	15.120	0.474	13.957	0.470	12.559	0.445
Mono-Energy	1.536	0.142	4.307	0.180	3.911	0.192	1.634	0.127	2.847	0.160
Stereo-Energy	1.511	0.139	4.301	0.180	3.895	0.191	1.612	0.124	2.830	0.159
INRAS [43]	1.405	0.141	3.511	0.182	3.421	0.201	1.502	0.130	2.460	0.164
NAF [26]	1.244	0.137	3.259	0.178	3.345	0.193	1.284	0.121	2.283	0.157
ViGAS [6]	1.049	0.132	2.502	0.161	2.600	0.187	1.169	0.121	1.830	0.150
AV-NeRF [21]	0.930	0.129	2.009	0.155	2.230	0.184	0.845	0.111	1.504	0.145
AV-GS [1]	0.861	<u>0.124</u>	<u>1.970</u>	<u>0.152</u>	**2.031**	<u>0.177</u>	0.791	<u>0.107</u>	1.417	<u>0.140</u>
SOAF [11]	<u>0.828</u>	0.126	**1.951**	0.153	2.097	0.182	**0.770**	0.109	<u>1.411</u>	0.142
π-AVAS (Ours)	**0.674**	**0.109**	1.992	**0.149**	<u>2.041</u>	**0.173**	<u>0.785</u>	**0.106**	**1.373**	**0.134**

Table 3. Quantitative comparison with state-of-the-art methods on the RAF dataset. We highlight the best-performing result in bold and underline the second-best result.

Methods	RAF-Furnished						RAF-Empty					
	T60↓	C50↓	EDT↓	Amp.↓	Phase↓	Env.↓	T60↓	C50↓	EDT↓	Amp.↓	Phase↓	Env.↓
AAC-nearest	13.0	3.41	73.5	1.09	1.60	4.83	13.0	3.41	73.3	1.09	1.60	4.83
AAC-linear	12.4	3.65	90.2	0.99	1.60	3.81	13.1	3.25	71.5	1.10	1.59	5.22
Opus-nearest	14.4	3.78	80.3	1.19	1.60	5.35	13.3	4.25	100.6	1.16	1.59	4.58
Opus-linear	13.1	3.55	77.8	1.47	1.60	5.74	12.7	3.94	95.5	0.95	1.59	4.26
NAF [26]	7.1	0.98	20.6	0.93	1.62	5.34	8.0	1.22	26.3	0.85	1.62	4.67
INRAS [43]	6.9	1.08	21.4	0.96	1.62	6.43	7.6	1.21	25.8	0.88	1.62	4.72
AVR [20]	<u>5.0</u>	<u>0.95</u>	<u>17.9</u>	<u>0.75</u>	<u>1.58</u>	**4.52**	<u>5.5</u>	<u>1.04</u>	<u>23.3</u>	<u>0.67</u>	<u>1.58</u>	**3.96**
π-AVAS (Ours)	**4.8**	**0.81**	**16.3**	**0.24**	**1.58**	<u>4.95</u>	**5.0**	**0.91**	**19.3**	**0.26**	**1.56**	<u>4.42</u>

We use MAG and ENV metrics to measure performance on the RWAVS dataset. Following AVR [20], we choose T60, C50, EDT, Amplitude (Amp.), Phase, and Envelope (Env.) to assess performance on the RAF dataset. T60, C50, and EDT are the most important metrics for measuring impulse response quality by analyzing energy decay. Amplitude and Phase metrics evaluate the impulse response in the time-frequency domain. The Envelope metric evaluates the impulse response in the time domain.

RWAVS Results. As shown in Tab. 2, we compare our method with existing baselines on the original RWAVS dataset. Mono-Mono, Mono-Energy, and Stereo-Energy are non-learnable methods that generate binaural audio by scaling mono audio using estimated energy values. Other methods are neural network-based approaches, such as AV-GS [1] and SOAF [11]. Our approach outperforms both non-learnable and learnable approaches, setting new state-of-the-art performance on the RWAVS dataset. The results demonstrate that our π-AVAS model achieves plausible novel-view audio synthesis quality.

RAF Results. We present our π-AVAS's performance on the RAF dataset in Tab. 3. AAC [16] and Opus [50] are traditional audio encoding methods. "Nearest" and "linear" refer to different interpolation modes. AVR [20] is the state-of-the-art method on this dataset. Our method surpasses AVR on most metrics and performs on par with it on the Envelope metric. Considering that AVR takes 24 hours to converge while our model trains in 5 hours, our π-AVAS exhibits a better trade-off between training time and impulse response generation quality.

4.3. Applicability Of Our Simulation Method

The generalization capability of our method is empowered by our first stage, the vision-guided audio simulation module (Sec. 3.1). We find that this module not only improves the performance of our flow matching model but can also be applied to existing neural synthesis approaches to enhance their generalization ability. By replacing their input source audio with our simulated audio, we integrate our vision-guided audio simulation module into their framework. Experiment results shown in Tab. 4 demonstrate the applicability of our approach, with overall performance improvements across most methods. For example, we improve the MAG metric of the AV-NeRF model by 0.772 and reduce the ENV loss by 0.036.

4.4. Ablation Studies

We provide a thorough ablation study using the RWAVS-Gen dataset, with results shown in Tab. 5.

Table 4. Applicability of our method. We apply our simulation method to other approaches to improve their generalization ability (denoted as "w/ sim"). We conduct experiments on the RWAVS-Gen dataset. The performance improvement is marked with a green triangle ▼.

Methods	Office		House		Apartment		Outdoors		Overall	
	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓
INRAS [43]	2.126	0.182	3.605	0.220	4.535	0.232	2.058	0.157	3.081	0.198
w/ sim	2.125	0.179	2.329	0.157	4.907	0.246	1.691	0.136	2.763 (▼ 0.318)	0.180 (▼ 0.018)
NAF [26]	2.275	0.181	2.873	0.186	4.878	0.231	1.575	0.135	2.900	0.183
w/ sim	2.203	0.184	2.214	0.154	4.925	0.241	1.556	0.131	2.724 (▼ 0.176)	0.178 (▼ 0.050)
ViGAS [6]	2.137	0.183	3.878	0.213	3.946	0.221	1.967	0.154	2.982	0.193
w/ sim	2.074	0.173	2.317	0.137	3.683	0.206	1.679	0.135	2.438 (▼ 0.544)	0.163 (▼ 0.030)
AV-NeRF [21]	2.086	0.180	3.759	0.221	4.520	0.230	2.308	0.165	3.168	0.199
w/ sim	2.014	0.174	1.946	0.136	4.374	0.221	1.250	0.122	2.396 (▼ 0.772)	0.163 (▼ 0.036)

Table 5. Ablation Studies. We conduct a comprehensive ablation study to verify the effectiveness of our proposed method. The term "pra+HRTF" refers to substituting our vision-guided acoustic simulation approach with `pyroomacoustics` and HRTF. "Regression" denotes training our audio refinement convolutional network without the flow matching loss.

Methods			Office		House		Apartment		Outdoors		Overall	
Simulation	Refinement	Augmentation	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓
pra[37]+HRTF[31]			4.259	0.226	5.264	0.227	8.762	0.288	3.454	0.203	5.435	0.236
✓			2.609	0.186	2.753	0.170	7.360	0.268	3.031	0.177	3.938	0.200
	Regression		1.904	0.167	3.826	0.204	4.033	0.219	1.612	0.143	2.844	0.183
	✓		1.880	0.166	3.144	0.198	4.209	**0.207**	1.577	0.138	2.702	0.177
✓	✓		**1.856**	**0.163**	2.191	0.148	**3.898**	0.209	1.496	0.134	2.360	0.164
✓	✓	✓	2.002	0.169	**1.946**	**0.140**	4.071	0.220	**1.326**	**0.125**	**2.336**	**0.164**

Vision-Guided Audio Simulation. First, we test the importance of our vision-guided audio simulation module (Sec. 3.1). We use `pyroomacoustics` [37] plus HRTF [31] as a baseline, which does not incorporate vision information. In this setup, we create a shoebox environment and use `pyroomacoustics` to estimate the room impulse response. We convolve the impulse response and the input audio to render mono audio at the target location. We then apply HRTF to generate binaural audio. Compared with this baseline, our module consistently outperforms it (see the first and the second rows), showing the importance of vision information in audio simulation.

Audio Refinement Network. We proceed to evaluate our second stage — the audio refinement flow matching model (see Sec. 3.2). To establish a baseline, we remove the flow matching training objective from the second stage and train the audio refinement network with a regression loss function, labeled as "Regression" in the table. By incorporating the flow matching training objective, we achieve more precise audio refinement performance (compare the third and fourth rows). We hypothesize that the flow matching formula decomposes the challenging one-step estimation into several simpler steps, thereby progressively refining the simulated sound. By combining our first and second stages (see the fifth row), we achieve improved performance beyond either stage alone, demonstrating (1) the realism limitations of simulation-only approaches, (2) the generalization challenges of neural rendering-only methods, and (3) the advantages of our two-stage approach.

Augmentation Strategy. We also assess the impact of our data augmentation strategy (refer to the last row). By enhancing the audio refinement training with additional data, we achieve lower metric losses for both house and outdoor scenes; however, we observe no improvement for office and apartment scenes. Consequently, we apply data augmentation only to house and outdoor scenes.

5. Conclusion

In this paper, we study the limitations of existing approaches to the audio-visual acoustic synthesis problem. We design a two-stage, physics-integrated audio-visual acoustic synthesis framework to enhance both realism and generalization capabilities. The first stage of our framework is a vision-guided audio simulation module, followed by a flow-matching-based audio refinement module. To mitigate data scarcity in this task, we also propose a data augmentation strategy. Experimental results show the effectiveness of our proposed approach, achieving new state-of-the-art results on the RWAVS-Gen, RWAVS, and RAF datasets. We further show how our physics-integrated method improves existing approaches in terms of generalization.

References

[1] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, Jiankang Deng, and Xiatian Zhu. Av-gs: Learning material and geometry aware priors for novel view acoustic synthesis. *arXiv preprint arXiv:2406.08920*, 2024. 2, 6, 7

[2] Amandine Brunetto, Sascha Hornauer, and Fabien Moutarde. Neraf: 3d scene infused neural radiance and acoustic fields. *arXiv preprint arXiv:2405.18213*, 2024. 2

[3] Chunxiao Cao, Zhong Ren, Carl Schissler, Dinesh Manocha, and Kun Zhou. Interactive sound propagation with bidirectional path tracing. *ACM TOG*, 35(6):1–11, 2016. 3

[4] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020. 2, 4

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

[6] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. *arXiv preprint arXiv:2301.08730*, 2023. 1, 2, 5, 6, 7, 8

[7] Mingfei Chen and Eli Shlizerman. Av-cloud: Spatial audio rendering through audio-visual cloud splatting. *Advances in Neural Information Processing Systems*, 37:141021–141044, 2025. 2

[8] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018. 2

[9] 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*, pages 21886–21896, 2024. 1, 2

[10] Roland W Fleming. Visual perception of materials and their properties. *Vision research*, 94:62–75, 2014. 4

[11] Huiyu Gao, Jiahao Ma, David Ahmedt-Aristizabal, Chuong Nguyen, and Miaomiao Liu. Soaf: Scene occlusion-aware neural acoustic field. *arXiv preprint arXiv:2407.02264*, 2024. 2, 6, 7

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

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[14] Chao Huang, Dejan Marković, Chenliang Xu, and Alexander Richard. Modeling and driving human body soundfields through acoustic primitives. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 2

[15] Chao Huang, Ruohan Gao, JMF Tsang, Jan Kurcius, Cagdas Bilen, Chenliang Xu, Anurag Kumar, and Sanjeel Parekh. Learning to highlight audio by watching movies. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23925–23935, 2025. 2

[16] International Organization for Standardization. Advanced audio coding (aac). ISO/IEC 13818-7:2006, 2006. 7

[17] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 3

[18] 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, 3

[19] Erwin Kreyszig. *Introductory functional analysis with applications*. John Wiley & Sons, 1991. 3

[20] Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. *NeurIPS*, 37:44600–44623, 2025. 2, 6, 7

[21] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. In *Thirty-seventh Conference on Neural Information Processing Systems*. 1, 2, 5, 6, 7, 8

[22] 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 arXiv:2309.15977*, 2023. 1

[23] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Language-guided joint audio-visual editing via one-shot adaptation. In *Proceedings of the Asian Conference on Computer Vision*, pages 1011–1027, 2024. 2

[24] 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. *arXiv preprint arXiv:2505.22865*, 2025. 2

[25] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*. 2

[26] Andrew Luo, Yilun Du, Michael J Tarr, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *NeurIPS*, 2022. 1, 2, 6, 7, 8

[27] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[28] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997. 2

[29] 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. In *ECCV*, 2020. 2, 3

[30] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31, 2018. 6

[31] Markus Noisternig, Alois Sontacchi, Thomas Musil, and Robert Holdrich. A 3d ambisonic based binaural sound reproduction system. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society, 2003. 8

[32] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2

[33] Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. Av-rir: Audio-visual room impulse response estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27164–27175, 2024. 2

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[35] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 6

[36] Lauri Savioja and U Peter Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 2015. 2

[37] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 351–355. IEEE, 2018. 8

[38] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 3

[39] Gabriel Schwartz and Ko Nishino. Recognizing material properties from images. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1981–1995, 2019. 4

[40] Samuel Siltanen, Tapio Lokki, Sami Kiminki, and Lauri Savioja. The room acoustic rendering equation. *The Journal of the Acoustical Society of America*, 122(3):1624–1635, 2007. 3

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*. 2

[42] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023. 6

[43] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. In *NeurIPS*, 2022. 1, 2, 6, 7, 8

[44] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 33:7537–7547, 2020. 4

[45] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023. 3

[46] Zhenyu Tang, Rohith Aralikatti, Anton Jeran Ratnarajah, and Dinesh Manocha. Gwa: A large high-quality acoustic dataset for audio processing. In *SIGGRAPH*, pages 1–9, 2022. 2, 4

[47] Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*. 2

[48] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12, 2016. 5

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2

[50] Xiph.Org Foundation. Xiph opus. https://opus-codec.org/, 2012. 7

[51] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021. 6

[52] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, pages 570–586, 2018. 2