COMP4801 Final Year Project

Interim Report

Geometry Aware Room Impulse Response Generation and Visualization

Yuemin Yu
UID: 3035945766
Supervisor: Professor Chenshu Wu
Date of Submission: 2025-01-26

Department of Computer Science,
School of Computing and Data Science,
The University of Hong Kong

Abstract

Room Impulse Response (RIR) describes the acoustic properties of a room, and it enables various applications such as sound source localization, Augmented Reality (AR) and Virtual Reality (VR). However, measuring RIR is challenging and time-consuming. Meanwhile, public datasets are limited in size, quality and diversity, and current available methods for RIR generation are still not performing satisfactorily. The visualization of the RIRs is also not provided in the current works, hindering the direct understanding of the generated or collected data. This project leverages the power of diffusion-based models to generate custom RIRs of superior quality conditioned on configurable room geometry and parameters, and provides 3D visualization using a Web application, to overcome the limitations. At present, public datasets have been collected for the training of the model, and experiments on currently available methods were conducted to serve as the baselines. The processing of the public data, the design and training of the model, as well as the evaluation and visualization of the generated data, are in progress. The expected outcome is a high-performance RIR generation and 3D visualization system, and potential future works will be explored to further enhance the system.

Contents

Abstract i

List of Figures iv

List of Tables v

List of Abbreviations vi

1 Introduction 1

1.1 Project Background . . . 1
1.2 Related Work and Motivation . . . 1
1.3 Project Objectives and Deliverables . . . 2
1.4 Outline of the Report . . . 2

2 Methodology 3

2.1 Dataset . . . 3
2.2 Model Design and Training . . . 3
2.2.1 Adoption of Diffusion Model . . . 3
2.2.2 Model Design and Training Strategy . . . 4
2.3 Evaluation . . . 4
2.4 System Development . . . 4
2.4.1 Visualization . . . 4
2.4.2 Web Application Development and Deployment . . . 5

3 Current Progress, Difficulties and Proposed Schedule 6

3.1 Current Progress . . . 6
3.1.1 Data Collection . . . 6
3.1.2 Model Design and Training . . . 6
3.1.3 Evaluation . . . 6
3.1.4 System Development . . . 6
3.2 Difficulties and Mitigation . . . 7
3.3 Proposed Schedule . . . 7

4 Future Work, Limitations and Conclusion 9

4.1 Future Work and Limitations . . . 9
4.1.1 Combination of Different Modalities . . . 9
4.1.2 Cloud Deployment . . . 9
4.1.3 Inference Acceleration and Edge Deployment . . . 9
4.2 Conclusion . . . 9

References **11**

List of Figures

1 Sound wave propagation in a room. 1
2 Image denoising and generation using DDPM. 3
3 2D visualization of WiFi signal. 5
4 RIR samples generated by different methods. 7
5 Edge devices with limited computing power. 10

List of Tables

1 Sizes of the collected datasets. 6
2 Proposed Schedule of the Project. 8

List of Abbreviations

RIR	Room Impulse Response
AR	Augmented Reality
VR	Virtual Reality
ML	Machine Learning
cGAN	Conditional Generative Adversarial Networks
DDPM	Denoising Diffusion Probabilistic Models
GPU	Graphical Processing Unit
REST	Representational State Transfer
API	Application Programming Interface
URL	Uniform Resource Locator
DOM	Document Object Model
AWS	Amazon Web Services
Azure	Microsoft Azure
GUI	Graphical User Interface
AGI	Artificial General Intelligence
HKU AIoT Lab	Artificial Intelligence of Things Lab, The University of Hong Kong

1 Introduction

1.1 Project Background

When sound signals travel within an indoor environment, absorption, reflection, diffraction, and attenuation might occur due to interaction with the walls, floor, ceiling and other obstacles, as illustrated in Fig. 1.

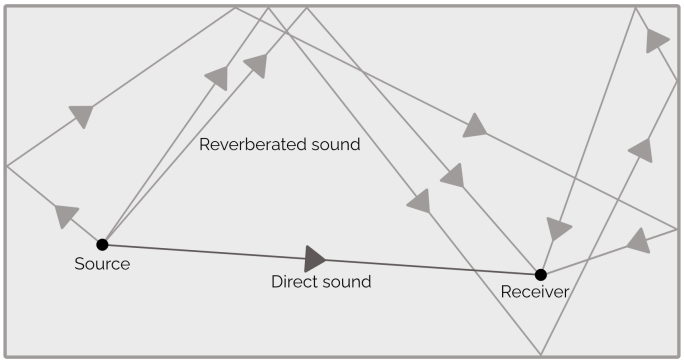


Figure 1: Sound wave propagation in a room. Different paths of the sound wave are produced because of the interaction with the room environment.

This phenomenon is represented by Room Impulse Response (RIR). RIR is a linear and causal time-domain filter [3], describing the influence of a given acoustic environment when a sound wave propagates from a source to a receiver.

RIR is a fundamental concept in acoustics and signal processing. It is widely used in different topics and tasks in acoustic field. For instance, when convolving clean speech with RIRs and adding background noise, a far-field speech training dataset can be synthesized [13], which can be used for downstream tasks. Other applications include acoustic events classification, sound source and receiver localization, AR, VR [12] and speech enhancement.

1.2 Related Work and Motivation

Measuring RIR often requires significant computational resources and lengthy processing times, making it challenging to deploy on edge devices with limited capabilities. Meanwhile, the public datasets of RIRs are very limited, making RIR generation an important task.

Several methods for generating RIRs have already been proposed previously. One of the methods is image-source method which was first proposed and developed in 1979 [1], and Diaz-Guerra et al. [4] demonstrated an GPU-based accelerated library for this method. Masztalski et al. [10] also proposed a stochastic method for RIR generation, which mainly serves for data augmentation tasks. In recent years, RIR generation using Machine Learning (ML) techniques has been explored with the development of generative ML models. One of the most famous works is the Conditional Generative Adversarial Networks (cGAN) based method [13], which

uses cGAN as the foundation model, with significant improvements in the quality of the generated RIRs as well as the runtime. Another work is an implementation of diffusion model [5], but the author claimed that the performance of the model is not as good as the cGAN-based method. The most recent works, [8] and [9], focus more on dereverberation and multichannel cases respectively.

Despite the progress, the performance of the current approaches still need to be improved. Gaps between the generated RIRs and the real ones still exist in various aspects, hence significantly reduce the value of the generated data in practical settings. In addition, previous works are not providing visualization of the output data samples as well, leading to difficulties for users to understand the generated data and to make rapid judgments. Therefore, a new system that can generate high-quality RIRs efficiently and visualize the output data samples is needed to address the gaps.

1.3 Project Objectives and Deliverables

The project attempts to develop an end-to-end system that can generate high-quality RIRs efficiently based a novel diffusion-based deep learning model, and visualize the output data samples to the users based on a Web application. The main objectives are following:

- **A novel diffusion model:** The project will develop a diffusion-based deep learning model for RIR generation task.
- **High-quality RIR generation:** High-quality RIRs will be generated based on the condition of input parameters collected with high accuracy in various evaluation metrics, and there will be significant improvements compared to the current available methods.
- **An end-to-end system with high-quality visualization:** Visualization will be rendered based on a Web application which can be deployed on different platforms, and several design strategies will be incorporated to ensure our system can run efficiently on different platforms.

1.4 Outline of the Report

The remaining part of the report is organized as follows: Section 2 introduces the project methodology, including the dataset, model design and training, evaluation, and system development. Section 3 presents the current progress of the project, the encountered or expected difficulties with their mitigations, as well as the proposed schedule. Section 4 discusses the limitations of the project with potential future works, and concludes the report.

2 Methodology

In this Section, the methodology of the project is introduced. Section 2.1 presents the sources of datasets. Section 2.2 explains the model design and training. Section 2.3 discusses the evaluation of the generated RIRs, and in Section 2.4, the system development is discussed.

2.1 Dataset

Public datasets in other works were collected for training and evaluation, and are being pre-processed and augmented. The datasets include the DECHORATE dataset [3], the MESHRIR dataset [7], and the RIR dataset from BUT ReverbDB [16]. The room geometry and configuration information contained in the datasets will be converted into embeddings and combined with the data samples to form the input of the model.

Meanwhile, these datasets are still in limited size, diversity and universality. The model will leverage the datasets for wider applicable range and configurations, enabling the output to be used in a broader variety of practical applications and scenarios.

2.2 Model Design and Training

2.2.1 Adoption of Diffusion Model

The project aims to utilize diffusion models to generate RIRs. Diffusion models are a class of generative models that have shown promising results in generative tasks in recent years. There are several famous diffusion model structures, for instance, Denoising Diffusion Probabilistic Models (DDPM) [6] and Generative modeling by estimating gradients of the data distribution [15]. Currently, most of the tasks conducted by such diffusion models are vision-based, including image inpainting, image generation (see Fig. 2) and other image and video processing tasks. On image generation tasks, the diffusion models outperform GANs [2], and they are also expected to have better scalability and stability than other models. Given the success of diffusion models in the image domain, the project attempts to further harness the strengths of these architectures in the audio domain.

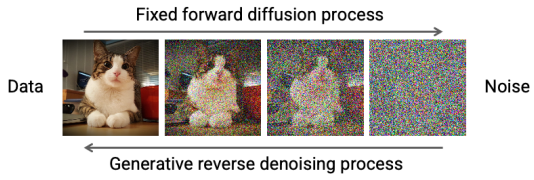


Figure 2: Image denoising and generation using DDPM. The Guassian Noise is added to the image, and the model is trained to remove the noise to generate the image.

2.2.2 Model Design and Training Strategy

The processed datasets with labels, which are described in Section 2.1, will be normalized according to the acoustic properties instead of the RGB values in the image domain. Moreover, hyperparameters will be tuned based the model performance. For instance, in the approximate posterior $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ of forward diffusion process [6], the value of variance schedule $\beta_1, \ldots, \beta_T$, as shown in Eq. 1, will be searched and optimized.

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) \coloneqq \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \qquad q(\mathbf{x}_t|\mathbf{x}_{t-1}) \coloneqq \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

While the primary loss is adapted from [14] for conditional generation tasks, as shown in Eq. 2, necessary modifications will be made to fit the RIR generation task due to the vast difference of data distribution between the image and audio datasets.

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right] \tag{2}$$

Other parameters including the learning rate, the optimizer, the batch size and the number of epochs, will also be configured for the best performance. PyTorch, which provides extensive support for the tuning of these parameters, is used as the main framework for the model implementation.

2.3 Evaluation

As mentioned in Section 1.3, the output will be compared with the current available methods as baselines with respect to a number of evaluation metrics, under the same condition embeddings. The evaluation metrics include but are not limited to the Reverberation Time Error τ_{60} and Word Error Rate in speech recognition tasks. The available baselines include the diffusion implementation [5], cGAN [13], stochastic method [10], and image-source method [4] mentioned in Section 1.2, and other methods available in the literature will also be explored. The generated RIRs will also be used in other downstream tasks, including speech enhancement and sound source localization, to perform head-to-head performance comparisons with the baselines and real RIRs.

2.4 System Development

2.4.1 Visualization

The visualization task is an important part of this project. Based on user input, RIRs will be generated and visualized interactively. A sample of 2D visualization of WiFi signal is shown in Fig. 3, and this project will utilize a similar approach, and 3D visualization technique will be applied to enhance the interactivity. The region of interest, room geometry, the position

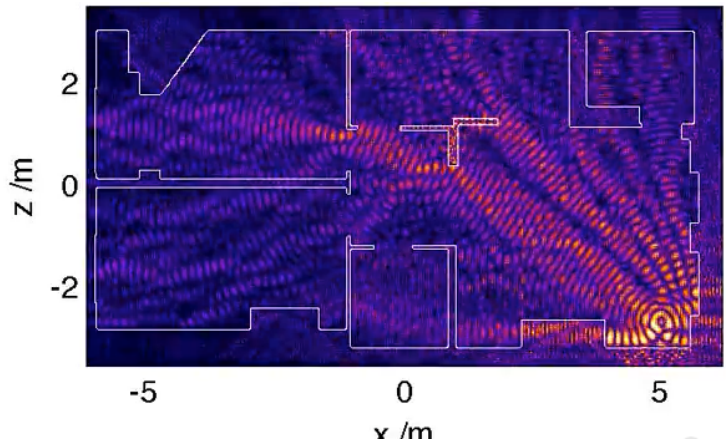


Figure 3: 2D visualization of WiFi signal. Signal strength is represented by color at different locations in the given room configuration.

of sound source and sound receiver from the user input will be rendered in 3D space, and the corresponding generated RIRs will be visualized in real-time, enabling users to make intuitive observations directly. The visualization will depend on a Web application, which will be discussed in the next section, Section 2.4.2.

2.4.2 Web Application Development and Deployment

The Web application will be developed using React.js for the frontend and Django for the backend. Upon completion of both the frontend and backend, the system will be deployed.

Frontend: Among all frontend frameworks and platforms available, React.js is chosen because of its simplicity, flexibility, stability and scalability in various Web-based applications. According to public data from StackOverflow [11], React.js is the most popular front-end framework in recent years, and many famous websites, such as Facebook, Instagram, and Netflix, are built on React.js, which makes resources and community support more accessible. React.js also operates serverside rendering and virtual Document Object Model (DOM), hence the load time can be reduced and better user experience can be provided. Moreover, React.js allows components reusing which greatly facilitates the development.

Backend: The backend is under construction using Django, a Web framework in Python. The major reason is that Django comes up with various built-in features, such as Representational State Transfer (REST)-ful Application Programming Interfaces (APIs) and Uniform Resource Locator (URL) routing, which can greatly accelerate the development of the backend. Moreover, since the backbone model is implemented in Python, Django can provide a seamless integration between the frontend and the backend.

Deployment: The system will be deployed and tested on end devices such as personal laptops to ensure the performance, efficiency and compatibility.

3 Current Progress, Difficulties and Proposed Schedule

3.1 Current Progress

3.1.1 Data Collection

The public datasets mentioned in Section 2.1 have been collected, and the size of each dataset is summarized in Table 1. The preprocessing and augmentation of the data are in progress due to the complexity of the datasets.

Dataset Name	**Size(GB)**
DECHORATE	83.9
MESHRIR	4.5
BUT ReverbDB	8.7

Table 1: Sizes of the collected datasets. DECHORATE is the largest and the most complex dataset, and augmentation might be necessary for MESHRIR.

3.1.2 Model Design and Training

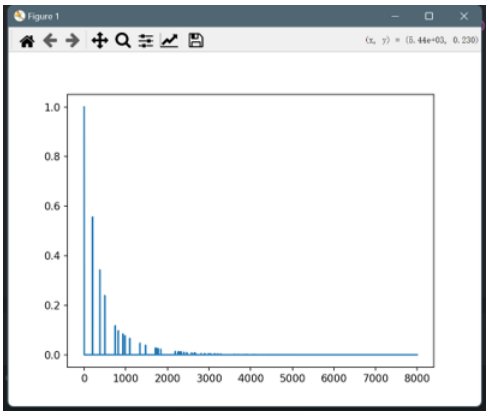
The initial network design and implementation have been completed. However, the parameter tuning and sweeping, the customization of the loss function and the training of the model are not yet finished, thus the model structure and parameters are subject to major changes.

3.1.3 Evaluation

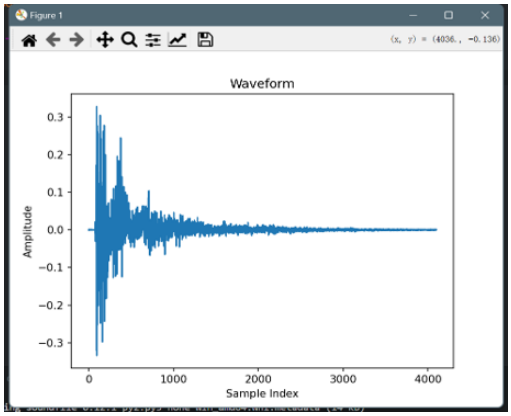
Comprehensive experiments were conducted on the current available baselines, including the cGAN-based method [13] and the stochastic method [10]. RIRs were generated using these approaches as shown in Fig. 4, and primary evaluation metrics were calculated. More experiments on other baselines, the design of evaluative tasks and exploration of more evaluation metrics are in progress.

3.1.4 System Development

The final rendering strategy of the visualization is still under research to keep the balance between the quality and the efficiency of the rendering. The design of Web application GUI has been completed and prototyped, yet the implementation of interactive frontend, backend and the design of APIs are still in progress.



(a) RIR from Stochastic Algorithm



(b) RIR from cGAN

Figure 4: RIR samples generated by different methods. The samples have different quality, length and sampling rate, therefore more evaluative tasks are needed for better comparison.

3.2 Difficulties and Mitigation

There are several difficulties encountered or expected in the project. The following are the major difficulties and the proposed mitigation strategies:

- The original parameters and loss functions of the diffusion model might not be suitable for the RIR generation task, and the search for the optimal combination need repeated experiments. Necessary mathematical calculations and proofs will be conducted based on the related works to derive some value bounds to guide the search.
- The training time of the model is long, and high-performance Graphical Processing Units (GPUs) are required. In order to reduce the training time, computational resources from the lab as well as cloud platforms are utilized.
- The primary evaluation metrics of the generated RIRs cannot effectively reflect the quality of the generated data, while the evaluative downstream tasks are complex and need extra time and resources to be conducted. Off-the-shelf software packages, libraries and tools will be fully adopted to facilitate the evaluation, which can save time and resources.

3.3 Proposed Schedule

The proposed schedule is shown in Table 2. In general, the progress is on track and aligns with the proposed schedule as I already finished the early stage of the project, and the next stage is in progress. The project is expected to be completed on time.

Date	Task	Current Status
2024.09 - 2024.11	Literature review on related topics of the project, prepare the dataset with large enough size for training, validation and testing.	Completed
2024.11 - 2025.03	Design and train the model, tune the model to get optimal output, develop the Web application.	In progress
2025.03	Test the system and model, and evaluate the performance of the project.	Planned
2025.04 - 2025.05	Update the project webpage accordingly, finalize the report and prepare for the final presentation with all deliverables ready.	Planned

Table 2: Proposed Schedule of the Project. The project is expected to be completed by May 2025.

4 Future Work, Limitations and Conclusion

4.1 Future Work and Limitations

Despite the expected wide range of applications and improved performance of the project, there still exists some limitations, with potential future works to be explored.

4.1.1 Combination of Different Modalities

Currently, the project only focuses on the RIR generation task without any visual and textural information, however integrating other modalities might further enhance the performance of the model. Some recent works provide multi-modality datasets which include RIRs and other information, such as [17]. The potential of using them as the input of the model or treat them as baseline will be explored. Moreover, the combination of RIRs, visual information and other types of signal can be studied to help real world perception of intelligent agents and systems, which eventually can contribute to the development of AGI.

4.1.2 Cloud Deployment

The system can be deployed on cloud platforms such as AWS, Azure and Google Cloud, which will further enhance the accessibility and scalability of the system.

4.1.3 Inference Acceleration and Edge Deployment

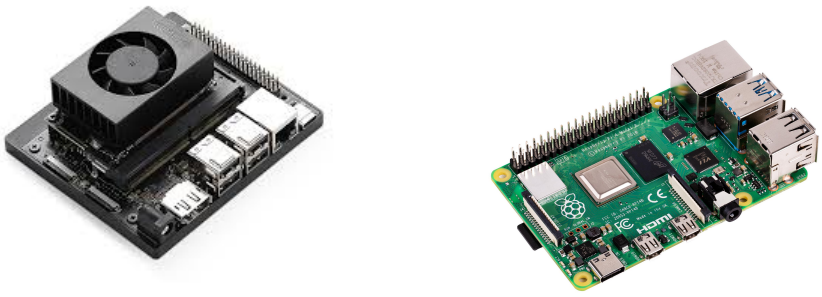
The inference speed of the model is crucial in ubiquitous and edge computing scenarios, while the current model might still suffer from long inference time and high requirements on computational resources. To further facilitate the ubiquitous computing tasks, modifications on the structure of the network can be explored to improve efficiency. If the inference speed can be improved to the level of seconds without the use of high-performance GPUs, then it can be deployed on edge devices such as Jeston Nano and Raspberry Pi, which have extremely limited computing power and simple structure without GPUs as shown in Fig. 5. This can further improve the accessibility of the system in ubiquitous computing tasks.

4.2 Conclusion

This project aim to develop a novel end-to-end RIR generation and visualization system. The generated RIRs are expected to have a significant boost in the performance compared to the current available methods including low τ_{60} Error and high performance in other evaluation metrics, and will be able to serve a wide range of practical applications, including but not limited to speech recognition, sound source localization, and AR/VR. In addition, the visualization will provide an intuitive understanding of the output RIR samples.

The early stage of the project has been completed, including the collection of datasets, the setup of baselines, and the design of the GUI of the Web application. The next stage, which is still in progress, focuses on the training of the model, the evaluation of the generated data, and the implementation of the system. The project is expected to be completed on time by May 2025.

Meanwhile, the limitations include the absence of multi-modality information, the potential long training and inference time, the high requirements on computational resources, and the lack of edge deployment. In the future, with further works on various aspects, this project can eventually contribute to wider applications with enhanced accessibility, including the development of intelligent systems, AGI, and ubiquitous computing.



(a) Jeston Nano **(b)** Raspberry Pi

Figure 5: Edge devices with limited computing power. They have simple but reliable structure and are widely used in edge computing scenarios.

References

[1] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, April 1979.

[2] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion Models Beat GANs on Image Synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[3] Diego Di Carlo, Pinchas Tandeitnik, Cedrić Foy, Nancy Bertin, Antoine Deleforge, and Sharon Gannot. DEchorate: a calibrated room impulse response dataset for echo-aware signal processing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):1–15, 2021.

[4] David Diaz-Guerra, Antonio Miguel, and Jose R. Beltran. gpuRIR: A python library for room impulse response simulation with GPU acceleration. *Multimedia Tools and Applications*, 80(4):5653–5671, October 2020.

[5] Eric Grinstein and Zehua Chen. RoomFuser: Room Impulse Response Generation using Neural Diffusion Models, October 2023.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2006.11239*, 2020. Accessed: 2024-09-23.

[7] Shoichi Koyama, Tomoya Nishida, Keisuke Kimura, Takumi Abe, Natsuki Ueno, and Jesper Brunnström. MESHRIR: A Dataset of Room Impulse Responses on Meshed Grid Points for Evaluating Sound Field Analysis and Synthesis Methods. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, 2021.

[8] Jean-Marie Lemercier, Eloi Moliner, Simon Welker, Vesa Välimäki, and Timo Gerkmann. Unsupervised Blind Joint Dereverberation and Room Acoustics Estimation with Diffusion Models. *arXiv preprint arXiv:2408.07472*, 2024.

[9] Francesc Lluís and Nils Meyer-Kahlen. Blind Spatial Impulse Response Generation from Separate Room- and Scene-Specific Information. *arXiv preprint arXiv:2409.14971*, 2024.

[10] Piotr Masztalski, Mateusz Matuszewski, Karol Piaskowski, and Michal Romaniuk. StoRIR: Stochastic Room Impulse Response Generation for Audio Data Augmentation. In *Interspeech 2020*, pages 2857–2861, 2020.

[11] Stack Overflow. Tag Trends of Frontend Frameworks. `https://trends.stackoverflow.co/?tags=reactjs,vue.js,angular,svelte,angularjs,vuejs3`, 2024. Accessed: 2024-09-23.

[12] 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 (CVPR)*, pages 27164–27175, June 2024.

[13] 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 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 571–575, 2022.

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, June 2022. IEEE Computer Society.

[15] Yang Song and Stefano Ermon. *Generative modeling by estimating gradients of the data distribution*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[16] Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Černocký. Building and evaluation of a real room impulse response dataset. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):863–876, 2019.

[17] Mason Long Wang, Samuel Clarke, Jui-Hsien Wang, Ruohan Gao, and Jiajun Wu. SoundCam: A Dataset for Finding Humans Using Room Acoustics. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.