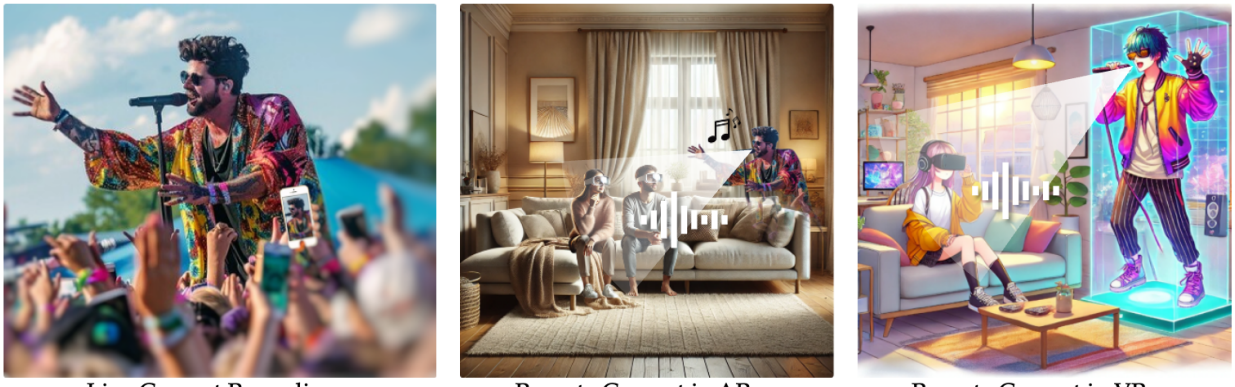
Multimodal Neural Acoustic Fields for Immersive Mixed Reality

Guansen Tong, Johnathan Chi-Ho Leung, Xi Peng, Haosheng Shi, Liujie Zheng, Shengze Wang, Arryn Carlos O'Brien, Ashley Paula-Ann Neall, Grace Fei, Martim Gaspar, and Praneeth Chakravarthula



Live Concert Recording | Remote Concert in AR | Remote Concert in VR

Fig. 1: We introduce a method to translate acoustic fields recorded from sparse viewpoints of a scene (*e.g.* live concert recording by multiple users) into remote and previously unseen scenes (*e.g.* remote concert in AR/VR) by leveraging diverse modalities including audio and visual cues, and spatial locations. Unlike existing methods that rely on complex 3D meshes of the environment for audio rendering, our framework uses audio and visual signals captured from discrete listener viewpoints. We employ a novel neural network to learn direct sound as well as the distinctive first-order reflections and multibounce reverberation patterns in a given scene from corresponding visual data, while also accounting for spatial acoustic variations. This allows for synthesizing realistic audio for unseen environments, both virtual and real-world.

Abstract—
We introduce multimodal neural acoustic fields for synthesizing spatial sound and enabling the creation of immersive auditory experiences from novel viewpoints and in completely unseen new environments, both virtual and real. Extending the concept of neural radiance fields to acoustics, we develop a neural network-based model that maps an environment's geometric and visual features to its audio characteristics. Specifically, we introduce a novel hybrid transformer-convolutional neural network to accomplish two core tasks: capturing the reverberation characteristics of a scene from audio-visual data, and generating spatial sound in an unseen new environment from signals recorded at sparse positions and orientations within the original scene. By learning to represent spatial acoustics in a given environment, our approach enables creation of realistic immersive auditory experiences, thereby enhancing the sense of presence in augmented and virtual reality applications. We validate the proposed approach on both synthetic and real-world visual-acoustic data and demonstrate that our method produces nonlinear acoustic effects such as reverberations, and improves spatial audio quality compared to existing methods. Furthermore, we also conduct subjective user studies and demonstrate that the proposed framework significantly improves audio perception in immersive mixed reality applications.

Index Terms—Multimodal interaction and perception, 3D user interfaces, Learning-based audio synthesis, Novel view synthesis, Acoustic spatialization, Augmented reality, Virtual reality

1 Introduction

Immersive audio rendering is crucial in augmented and virtual reality (AR/VR) as it enhances the sense of realism and spatial awareness by accurately simulating how sounds originate and move in a 3D environment . For example, when a user attends a remote concert in AR/VR, adjusting the audio direction, distance, and reverberation of the sound to match the visual cues of the user's physical or virtual environment creates a more immersive experience, as illustrated in Figure 1. As the listener moves around, the system should dynamically adjust how they hear the music – whether they are closer to the stage, standing under a balcony, or hearing the distant echoes from the crowd – making the experience highly immersive and realistic. This realism, created by accurate spatial audio rendering, transforms the concert from just watching a performance into a fully engaging, lifelike experience. Similarly, in a virtual meeting, if someone speaks from a position behind you, ensuring that the sound arrives as it would in real life, from behind, enhances spatial cognition.

Existing methods for immersive audio face several challenges in accurately capturing the complexity of sound propagation and its interaction within a given space, limiting their effectiveness. Popular techniques such as binaural rendering or standard spatial audio often rely on simplified models that inadequately simulate how sound reflects, diffracts, and is absorbed by the environment [25, 61]. Especially, these methods struggle with real-time performance, particularly in complex environments such as concert halls, where multiple sound sources and dynamic interactions occur simultaneously. This leads to unrealistic reverberations, echo effects, and a lack of spatial depth, breaking immersion [49, 61]. Moreover, existing approaches typically use generalized Head-Related Transfer Functions (HRTFs), which do not account for variations in sound perception due to differences in individual head and ear shapes. Without personalized HRTFs, the spatial accuracy of the audio can be compromised, diminishing the overall quality of immersion [15, 39, 67]. Additionally, the computational demands of

• *All authors are with the Department of Computer Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

rendering high-fidelity immersive audio in real time, especially for complex multi-source environments, can lead to latency issues and reduced audio quality, further limiting the effectiveness of current methods [50].

Recent novel-view synthesis and 3D reconstruction methods such as neural radiance fields (NeRF) [38] and Gaussian splatting [22], have demonstrated great success in recovering a complex scene from discretely captured image data. These techniques have also demonstrated significant benefit in terms of compute and rendering for AR/VR edge-device applications [8, 21]. However, such approaches do not directly apply for spatial audio rendering as sound interacts with environments in more complex ways than light. Moreover, handling the multi-source and time-sensitive nature of sound adds additional complexity in applying Neural Radiance Fields (NeRFs), originally developed for rendering realistic 3D visual scenes, for rending audio in real-time without compromising the audio quality or accuracy. This challenge is further compounded if sound from a given environment recorded at a sparse set of locations needs to be translated and adapted to a completely novel unseen 3D environment, with appropriate spatial audio effects.

In this work, we introduce Multimodal Neural Acoustic Fields (MNAF) framework to enable immersive audio in AR/VR, overcoming some of the challenges described above. Specifically, our method consists of two modules: 1) a visual-acoustic fusion module that uses a hybrid transformer-convolutional neural network, dubbed conformer, to learn the mapping between visual features of the ambient scene images (obtained from a pre-trained visual feature extractor) and input source audio signals, and 2) a waveform synthesis module that takes fused audio-visual features from the previous module to synthesize audio as heard from a novel unseen viewpoint. The framework is trained on a large dataset of diverse environments and audio signals, and once trained, it allows for generating audio from novel viewpoints. Our framework also allows for translating the source audio into a completely unseen 3D environment and incorporating personalized HRTFs to synthesize binaural audio for individual users. We validate our approach on synthetic and real audio-visual data and showcase visible improvements over prior works in sound synthesis and rendering room acoustic effects. Subjective user studies indicate that our framework provides perceptually enhanced audio experience.

In summary, the contributions of our paper are as follows:

- We propose a novel learning-based framework to synthesize spatial audio from unseen viewpoints and environments, leveraging visual-acoustic features and incorporating a multimodal conformer with an adaptive convolutional module featuring learnable kernel dilations and positions.
- We validate our framework on both synthetic and real-world datasets, as well as a custom dataset of conversational content recorded in diverse environments with varying noise levels, and demonstrate our framework's effectiveness across a range of scenarios.
- We conduct a series of subjective user evaluations of our framework using everyday scenarios rendered in virtual reality to assess the quality of immersive audio rendering at unseen viewpoints, and demonstrate enhanced immersive experience.

Datasets, training code, trained model weights and additional material will be made publicly available.

Overview of Limitations: We take a first step towards rendering immersive audio at unseen locations and environments via multimodal audio-visual data. While our proposed approach demonstrates clear improvements in generating immersive audio signals from new viewpoints, refining the method to directly support 360-degree video inputs will enhance immersion in VR and AR applications. Incorporating additional modalities that complement visual data [37] might further improve the quality of immersive experiences. Furthermore, extending the method to explicitly consider the effect of materials in the scene on sound attenuation, and rendering audio in real-time is an exciting future direction.

2 Background and Related Work

Our work closely relates to learning-based audio synthesis, novel view synthesis, and acoustic spatialization, which we discuss here.

2.1 Learning-Based Audio Synthesis

Traditional audio processing systems synthesize sounds by assembling audio segments or using task-specific mathematical models [51] tailored to specific applications. Recent advances in deep neural networks have significantly improved audio synthesis, demonstrating remarkable capabilities in generating realistic voices and music notes [64] and matching room acoustics [5], despite the challenges posed by audio's highly dynamic and time-sensitive nature. Autoregressive models such as WaveNet [64] and SampleRNN [36, 72] can generate high-quality audio and switch speakers for text-to-speech synthesis, but they struggle to maintain temporal consistency over long sequences, leading to potential degradation in audio quality. Adversarial Audio Synthesis [9, 10] addresses this issue using global latent conditioning and parallel sampling while relying heavily on the discriminative model for quality. Diffusion models [24, 44, 73] have emerged as a promising approach, converting white noise into structured waveforms with fast inference and generalization. Yet, they face challenges in model size, long-term coherence, and dependence on large training data. Our work enhances existing models and methods by improving temporal consistency and audio quality through audio-visual learning. By incorporating visual features, we refine the synthesis process, enabling more accurate rendering of effects such as reverberation and echoes.

2.2 Novel-View Synthesis

Recent advances in neural radiance fields (NeRF) [38, 55] have revolutionized novel-view synthesis (NVS) by learning continuous and implicit scene representations using multi-layer perceptions (MLP). The original NeRF approach relies on end-to-end training of volume rendering models to represent static scenes, requiring a large number of calibrated images from multiple viewpoints for high-quality results [38]. Subsequent works have explored various extensions, including reducing the number of input views [18, 26, 41, 47, 70] and synthesizing dynamic scenes [8, 29, 32, 43, 45, 60, 63]. More recently, 3D Gaussian splatting has emerged as a promising approach for unbounded and complete scene representation in NVS [20, 22, 46]. These advancements in visual scene representation inspire research into the representation of sound in the acoustics domain.

2.3 Acoustic Spatialization

Acoustic spatialization aims to create realistic auditory experiences by modeling how sound propagates and interacts within three-dimensional environments [6]. Traditional methods often rely on computationally expensive or scene-specific approaches, such as ray-tracing-based simulation [25, 50, 54, 65] or wave-based simulations [33, 62] by empirically estimating acoustic properties [3, 27]. Therefore, it is challenging to apply these approaches to arbitrary scenes.

Recent works explored neural networks to learn implicit representations that can continuously map spatial coordinates of a scene to corresponding audio, enabling flexible and compact modeling of the acoustic field. Acoustic Scattering Fields, Neural Acoustic Fields (NAF), INRAS [35, 59, 61], and IR-MLP [48] represent seminal works in this domain. These models utilize a framework of MLP that takes listener and source positions as input, and outputs the waveform of the impulse response, a rendition of audio in the time domain. These approaches showed success in conditioning the network on local geometric information extracted from a trainable feature grid, allowing for the disentanglement of geometrical features into arbitrary emitter and listener locations within a scene. However, they rely on 3D meshes and are limited in their application to a single source noiseless environment, making them impractical for immersive AR/VR applications. In contrast, our model relies on captured images from the listener's viewpoint and relative pose to the speaker, all of which will be freely available from the sensors on wearable display eyeglasses, allowing the method to generalize effectively across diverse environments.

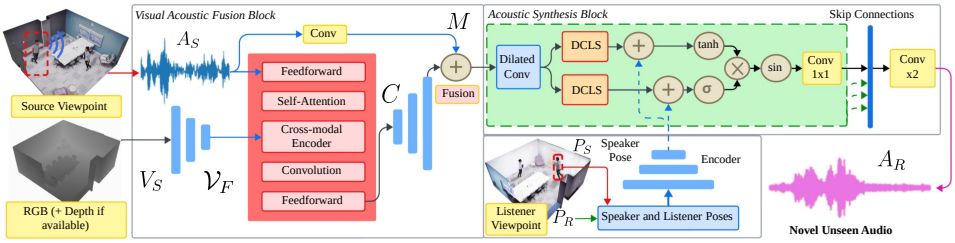


Fig. 2: **Model Architecture**. Given source audio input, source RGB image and optionally depth, we first encode the RGB channel data through a Conformer which produces C at the current view point. We then upsample respectively this reverberated audio and the source audio and concatenate them in the latent space. We then pass this the into a Waveform Synthesis block with adaptive convolutions. We use positional features at the novel viewpoint as conditions for the waveform synthesis block. The positional features are obtained by concatenating positional information from the listener viewpoint, namely the listener pose and rotation, as well as the speaker pose, then learning a single representation through the MLP. In each block, the audio sequence is processed by adaptive conv1d layers and the positional features are processed by conv1d layers.

AV-NeRF [31], Visual Acoustic Matching [6], and NVAS [5] utilized an end-to-end neural network to fuse localized visual features of the environment with acoustic features. However, these methods fall short of capturing the complex relationships between the visual environment and its acoustic characteristics without leveraging the temporal correlations between audio and visual features. In addition, the binaural channels of the stereo audio are often trained separately in these models, limiting effective learning of the spatial acoustics, which compromises immersive experiences. In contrast, our method extracts acoustically relevant information from visual features by directly learning the audio-visual correlation in the temporal domain. With effective cross-modal fusion, our proposed framework enables perceptually high quality sound rendering, as also validated by our perceptual studies.

3 Novel-View Acoustic Synthesis

We seek to synthesize audio as heard from a novel viewpoint of a given environment or in an unseen new environment using the audio-visual sensory measurements on an AR/VR headset. This task is intrinsically multi-modal and challenging. Our method takes as input audio-visual data (A_S, V_S) captured from a "source" (or speaker) viewpoint, assuming the position and orientation of the speaker relative to the "receiver" (or listener) viewpoint are known, as illustrated in Figure 2. Using this information, we synthesize the audio A_R at a novel receiver viewpoint for any scene via a learnable mapping,

$$f : (A_S, V_S, P_S, P_R; E) \mapsto A_R, \tag{1}$$

where P_S, P_R describes the speaker and listener poses, respectively, for a given environment E.

Our proposed multimodal neural acoustic fields framework learns the transfer function $A_R = f_{E_R}(A_S, V_S, P_S, P_R; E_S)$ to synthesize audio at novel receiver positions P_R and environments E_R given the input audio in the source environment E_S. The acoustic characteristics of the scene, however, are not explicitly handled but are implicitly learnt from the audio-visual data. The higher-order acoustic effects such as reverberations depend on the geometric features of the environment whereas direct line-of-sight sound depend on the relative poses of the speaker and listener. Therefore, we infer A_R in two stages: the higher-order reflections from the audio-visual features A_S and V_S, and direct sound from the poses P_S and P_R, as shown below:

$$A_R = f_{\text{AS}}\left(f_{\text{VAF}_{E_R}}(A_S, V_S; E_S), P_S, P_R; E_R\right) \tag{2}$$

where f_{VAF} denotes the first visual-acoustic fusion stage of inferring reverberations in the new environment based on the sampled sound and 3D space of the source environment, and f_{AS} denotes the second acoustic synthesis stage where the direct sound is inferred based on the relative poses of the speaker and listener. Next, we describe our two stage method, together dubbed as Multimodal Neural Acoustic Fields.

4 Multimodal Neural Acoustic Fields

Our Multimodal Neural Acoustic Fields (MNAF) frameworks consists of two core components: the **Visual-Acoustic Fusion Block** and the **Acoustic Synthesis Block**, as discussed in Section 3. The Visual-Acoustic Fusion block is built as a hybrid transformer-convolutional neural network, also known as conformer, to implicitly learn the scene acoustics. It has been shown that such a model can also implicitly learn the material properties such as the acoustic impedance and surface roughness which influence higher-order reverberations, purely from visual features [34]. We specifically use a cross-modal encoder to model the correlation between visual data and corresponding audio, and encode scene-dependent acoustic characteristics into latent features using our visual-acoustic fusion conformer network. The encoded latent features are then passed to an acoustic synthesis block to decode the acoustic field as well as synthesize direct sound and inter-aural effects dependent on the relative location and pose of the speaker and the listener. This section discusses our framework in detail.

4.1 Visual-Acoustic Fusion for Encoding

The Visual-Acoustic Fusion (VAF) block employs a stacked conformer architecture [5, 16] to capture the effects on reverberations informed by visual data. Simulating accurate acoustic material properties for every object is challenging in real-world scenarios. Therefore, we rely on visual features to serve as proxies to object material properties [2, 12, 28, 40, 53], translating learned object characteristics into their corresponding sound effects. For instance, hard surfaces like metals result in sharper and intense reflections with longer reverberation times, while soft materials like fabric absorb sound, dampening reverberation. Each conformer block consists of a feed-forward module, a Multi-Head Self-Attention (MHSA) module, a cross-modal encoder for cross-attention, a convolutional module, followed by another feed-forward module in sequential order, as illustrated in Figure 2. Before applying cross-modal attention, we encode the RGB(D) visual features of the source viewpoint V_S as $\mathscr{V}_F$ by a pretrained ResNet18 network [17]. The cross-attention is inserted after the self-attention to establish the correlation between video patches of the source view and acoustic characteristics. The attention score $\mathscr{F}_{cm}$ between the pairs of audio A_S and visual features $\mathscr{V}_F$ are computed as

$$\mathscr{F}_{\text{cm}}(A_S, \mathscr{V}_F) = \text{softmax}\left(\frac{A_S \mathscr{V}_F^T}{\sqrt{H}}\right)\mathscr{V}_F \tag{3}$$

where H is the feature dimension of A_S and $\mathscr{V}_F$.

The cross-modal attention establishes correlations between the audio-visual inputs of the source view. This approach effectively draws inferences from their corresponding spatial and material-dependent room acoustic characteristics from audio-visual data. As a result, the visual-acoustic fusion block learns an implicit representation $\mathscr{C}$ of room acoustic characteristics in the temporal domain and is encoded as latent features of the scene-dependent visual-acoustic data:

$$\mathscr{C} = \text{Conformer}(A_S, V_S). \tag{4}$$

While the conformer captures the higher-order reflections and reverberation within the scene, decoding and synthesizing the audio at a novel viewpoint requires the audio from the source view. Therefore, we pass the latent vector $\mathscr{C}$ and the source audio A_S through 1D convolutional layers for resampling and matching their dimensions before they are fused through an additional fusion layer, to obtain output

$$M = \text{Fusion}(\hat{C}, \hat{A_S}) \tag{5}$$

where $\hat{C}$ and $\hat{A_S}$ are the result of 1D convolutional resampling.

4.2 Acoustic Synthesis Network for Decoding

The acoustic synthesis block decodes the conformer output from the previous stage and uses the listener's relative position information from the speaker (in the new environment) to conditionally generate spatial audio. Positional inputs such as the relative pose of the source and receiver plays a crucial role in perceiving direct sound. Therefore, the relative positions and orientations of the speaker P_S and listener P_R are passed through sinusoidal positional encoding to get an embedding of the spatial coordinates, $\gamma(P_S)$ and $\gamma(P_R)$, respectively.

The encoded spatial information is then transformed through a series of linear projections via an MLP to map the positional embeddings into a lower dimensional positional feature space, denoted as

$$\hat{z} = \text{MLP}\big(\gamma(P_S), \gamma(P_R)\big).$$

These features enable our model to capture the spatial dynamics in the scene such as the effect of the listener's motion on the perception of spatial sound, and are passed to the acoustic synthesis module as conditional input. The acoustic synthesis module then decodes the output from visual-acoustic fusion block M (Equation (5)) by conditioning on positional features $\hat{z}$. The acoustic synthesis decoder is designed as a network of N stacked synthesis blocks as shown in Figure 3, where each block consists of multiple 1D convolutional layers. Every block features gated adaptive layers to dynamically adjust and learn the influence of positional data on the encoded audio. Specifically, this design ensures that the model accurately captures how spatial changes such as distance and orientation affect the perceived sound. More details on adaptive convolution is discussed in Section 4.3.

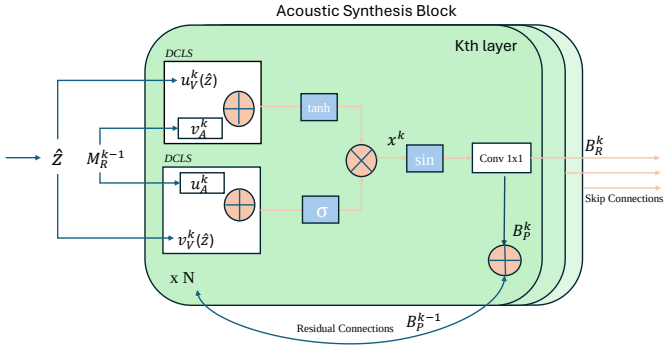


Fig. 3: **Acoustic Synthesis Block.** We decode the fused audio-visual features M from the previous stage by conditioning on positional feature space of the speaker and listener $\hat{z}$ to generate spatial audio.

The VAF block provides fused audio-visual features which may or maynot be relevant to current listener's position. To distill relevant information, we regulate the complex variations in audio-visual data and extract the features relevant to current listener's pose. To this end, we use a tanh activation on M, and a sigmoid activation on $\hat{z}$ as a gate to regulate the flow of information. Specifically, we employ a Hadamard product of tanh and sigmoid activation functions with learnable weights to serve as the filter and gate respectively, and is described by

$$x^k = \tanh\left(u_A^k(M_R^{k-1}) + u_V^k(\hat{z})\right) \odot \sigma\left(v_A^k(M_R^{k-1}) + v_V^k(\hat{z})\right), \tag{6}$$

$$B_R^k = w_1\left(\sin x^k\right), \quad B_P^k = w_2\left(\sin x^k\right), \tag{7}$$

where $k = 1,...,N$ is the layer index, x^k is the intermediate feature representation at the k-th layer, $\tanh(\cdot)$ and $\sigma(\cdot)$ are the activation functions, and $\odot$ is the Hadamard product. The learnable weights for encoded audio and video features in the k-th adaptive convolutional layer are denoted by $u_A^k, u_V^k, v_A^k, v_V^k$. As illustrated in Figure 3, the k-th layer of the acoustic synthesis module takes values from the residue connection of the previous (k-1)-th layer, B_R^{k-1}, and outputs the features B_P^k that are fed to the next layer. The result of these gated networks are passed through a sinusoidal activation function, and encoded with two convolutional weights w_1 and w_2 (Equation (7)). The sinusoidal function introduces periodicity and maintains a uniform amplitude, making it well-suited for audio encoding [6]. The outputs $\{B_P^k, (k = 1,\ldots,N)\}$ are all processed by mean pooling before finally decoding to produce the synthesized audio $\hat{A}_R$.

4.3 Adaptive Convolutions in Acoustic Synthesis Block

Modeling long-term dependencies in audio requires a large receptive field to capture the necessary long-range information. One approach to this challenge is to stack 1D convolutional layers with exponentially increasing dilation of spacings. While dilated convolutions expand the receptive field, their fixed spacings limit the model's ability to capture specific frequency patterns in audio data. As dilation increases across layers, large gaps between sampled points cause the network to miss subtle, localized changes in the audio sequence, due to which the network struggles to learn the fine-grained waveform variations.

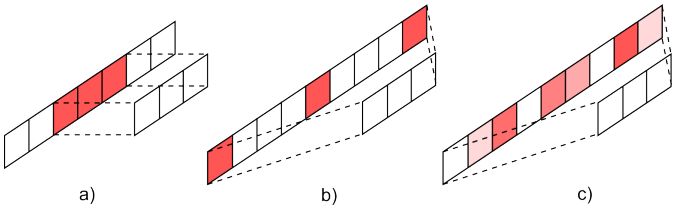


Fig. 5: **Dilated Convolution with Learned Spacings (DCLS).** (a) a standard 1D convolution with kernel size of 3. (b) a dilated 3 kernel with dilation rate 2. (c) a 1D convolution with learned dilation spacings with 3 kernel elements and a dilated kernel size of 9.

We overcome this in our acoustic synthesis module by integrating dilated convolutions with learnable spacing (DCLS) [23] into the waveform synthesis network. This allows our model to dynamically adjust and optimize the dilations, thereby effectively capturing both long-term dependencies and fine-grained details presented in the audio signals. We show in ablation studies that this approach leads to overall better performance, especially in unseen data and environments.

A DCLS module with a kernel K of m elements is formulated as

$$F : \mathbf{w}, \mathbf{p} \mapsto K = \sum_{i=1}^{m} f(w_i, p_i), \tag{8}$$

where i ($1 \le i \le m$) denotes kernel elements, w_i denotes their learnable weights, p_i, $1 \le p_i \le s$ are the learnable positions along the kernel, and $f(w, p)$ is the contribution from each kernel element. As opposed to a regular kernel (Figure 5a), a kernel with learned dilation spacing (Figure 5b,c) can have kernel elements in arbitrary locations, thereby extending the overall receptive field. However, since the kernel positions are also learned, they can be fractional.

The final kernel elements are therefore computed via interpolation of elements of the vector K with learned parameters w, p as

$$K_\ell = \begin{cases} w(1-r) & \text{if } \ell = \lfloor p \rfloor \\ wr & \text{if } \ell = \lfloor p \rfloor + 1 \end{cases} \tag{9}$$

and $r = p - \lfloor p \rfloor$ is the fractional part of p, and overlapping kernel weights at a given position are added.

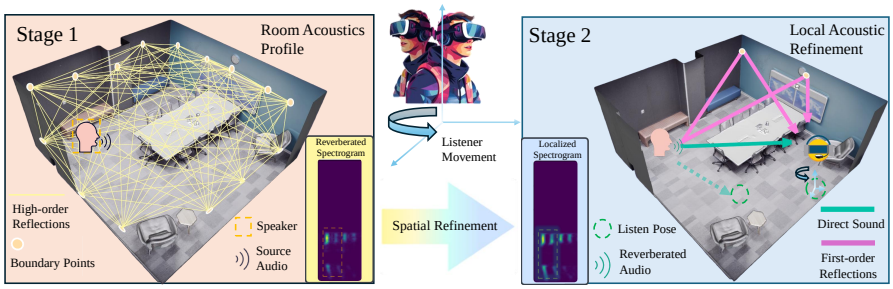


Fig. 4: **Multimodal Neural Acoustic Field (MNAF).** Sound propagation can be classified into direct sound emission, early reflections, and higher-order reflections, also known as reverberations. In our model, we address sound propagation in two stages. First, we model reverberations based on the room's consistent acoustic profile, which is determined by its geometry, materials, and other characteristics learned from visual features. Then, we refine this model by incorporating how direct sounds and early reflections are influenced by the pose of the speaker and listener. In the encoding phase, described in Section 4.1, we extract visual features to model the room's reverberant sound field, capturing the overall acoustic profile of the environment. This profile remains consistent throughout the room. In the next step, we incorporate local acoustic effects, which change as the speaker or listener moves within the space. Detailed in Section 4.2, we utilize positional data, such as the speaker's pose and orientation, to further refine direct and early reflective sound paths. To achieve precise spatial audio effects, we also optimize interaural differences using a stereo loss function. Additionally, our proposed HRTF module and its preliminary validation are discussed in Section 6.3.

4.4 Loss Function

Our loss function includes an ℓ_1 loss between the magnitude of short-time Fourier transform (STFT) of predicted audio $\hat{A_R}$ and groundtruth audio $\hat{A_R}$,

$$L_{\text{mag}} = \big|\,|\text{STFT}(\hat{A}_R)| - |\text{STFT}(A)|\,\big|. \tag{10}$$

One noted issue with only using a STFT magnitude loss is that the model may fail to distinguish energy differences across the binaural channels, leading to short-cut learning [13, 57]. Therefore, to help the model better learn stereo balance, we compute the sum and difference signals on two channels,

$$A_{R,\text{sum}} = A_{R,\text{left}} + A_{R,\text{right}} \tag{11}$$

$$A_{R,\text{diff}} = A_{R,\text{left}} - A_{R,\text{right}} \tag{12}$$

where $A_{R,\text{sum}}$ captures the shared audio content between channels (*i.e.* mono sounds evenly distributed across both speakers), and $A_{R,\text{diff}}$ captures the differences, encoding stereo separation and spatial details such as one channel being louder or delayed relative to the other.

We also employ an additional multi-resolution loss ℓ_{MR} [69] composed of spectral convergence ℓ_{SC} and spectral log magnitude ℓ_{SM} to captures both fine and coarse spectral features:

$$\ell_{\text{SC}}(\hat{A}_R, A_R) = \frac{\big\|\,|\text{STFT}(A_R)| - |\text{STFT}(\hat{A}_R)|\,\big\|_F}{\big\|\,|\text{STFT}(A_R)|\,\big\|_F} \tag{13}$$

$$\ell_{\text{SM}}(\hat{A}_R, A_R) = \frac{1}{N}\,\big\|\log\left(|\text{STFT}(A_R)|\right) - \log\left(|\text{STFT}(\hat{A}_R)|\right)\big\| \tag{14}$$

$$\ell_{\text{MR}}(\hat{A}_R, A_R) = \frac{1}{M}\sum_{m=1}^{M}\left(\ell_{\text{SC}}(\hat{A}_R, A_R) + \ell_{\text{SM}}(\hat{A}_R, A_R)\right), \tag{15}$$

where ℓ_{sc} supervises the deviation of STFT amplitude from ground truth and ℓ_{sm}, the log amplitude of STFT, provides additional quantization levels for learning weaker signals. We apply ℓ_{MR} to each channel as follows,

$$L_{\text{stereo}} = \ell_{\text{MR}}(\hat{A}_{R,\text{sum}}, A_{R,\text{sum}}) + \ell_{\text{MR}}(\hat{A}_{R,\text{diff}}, A_{R,\text{diff}}) \tag{16}$$

Our final optimization objective is defined as

$$L_{\text{total\_loss}} = L_{\text{mag}} + \alpha L_{\text{stereo}} \tag{17}$$

where α is a weight parameter.

We implemented the proposed framework in PyTorch and all the models were trained on a single NVIDIA RTX 3090 GPU. We used the ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and $\alpha = 0.02$ for training, with each model trained for 10 epochs with a learning rate of 5×10^{-4}.

5 Experiments and Evaluations

5.1 Simulation and Real-World Datasets

We used three data sets for our model evaluation: **Replay** [52], **SoundSpaces** [6], and a custom collected data set which we call **EnvSound**. The **Replay** data set contains multiview audio and video recordings of real-world conversational scenarios captured in a home-like environment. Eight stationary cameras and binaural microphones simultaneously record 5-minute conversations split into videos of one second each, providing synchronized multiview frames and audio. For training, we randomly selected two out of the eight viewpoints (yielding 56 possible speaker-listener combinations per scene), with the dataset split into 77K/12K/2K clips for train/validation/test. The visual input resolution is 256×256.

The **SoundSpaces dataset** simulates room acoustics based on Matterport3D [4] and Replica [58] 3D environments, offering 120 virtual scenes with 1,000 speakers and 200,000 viewpoints. For each environment, we randomly sampled two speaker locations and four nearby viewpoints oriented toward the speakers and recorded a one-second video. Each speaker was assigned a gender-matched speech sample from the LibriSpeech dataset [42], and binaural impulse responses were computed for all speaker-viewpoint pairs. During training, two of the four viewpoints were randomly selected as source and target for each sample. The SoundSpaces dataset was split into two testing categories: "Same Environment" (testing on environments seen during training) and "Novel Environment" (testing on unseen Gibson environments), using a 90/10/20 split for train/val/test. The visual input resolution is 216×384.

To assess the robustness of our model, we created a custom multi-environment, multi-scenario dataset, which we name **EnvSound**, as SoundSpaces lacks ambient noise and Replay is limited to a single environment. We evaluated our model's performance on this dataset using both objective metrics and user studies. Our custom dataset comprises of video clips ($\sim$ 30s) recorded simultaneously using two iPhone 14s (24mm focal length, 1080x1920 resolution at 30 fps, MEMS stereo microphones) in diverse, uncurated everyday settings (15 scenes in total), with the phone held in random positions relative to the speaker. It features a variety of scenarios, such as conversations and monologues, and offers a broader range of noise levels compared to existing datasets. For instance, some clips, captured in noisy environments like busy restaurants, feature ambient sounds comparable in volume to the conversations, while others, recorded in quiet spaces like offices, have minimal background noise. The dataset also spans a wide variety of visual settings, including different room sizes, lighting conditions, and levels of visual clutter. No denoising or speech enhancement was applied to the recordings, allowing us to rigorously test the model's performance under challenging noise conditions.

	Synthetic Dataset						Real-world Dataset					
	Same Environment			Novel Environment			Replay			EnvSound		
	Mag↓	LRE↓	RTE↓	Mag↓	LRE↓	RTE↓	Mag↓	LRE↓	RTE↓	Mag↓	LRE↓	RTE↓
Input Audio	0.225	1.473	0.032	0.216	1.408	0.039	0.153	1.322	0.045	0.043	5.52	0.042
TF Estimator	0.648	2.713	0.066	0.815	2.792	0.067	0.334	2.632	0.142	-	-	-
DSP	0.634	7.194	0.049	0.876	7.775	0.048	-	-	-	-	-	-
ViGAS	0.150	1.260	0.037	0.211	1.241	0.036	0.145	0.947	0.049	0.043	5.25	0.043
Ours	**0.134**	**1.151**	**0.034**	**0.186**	**1.125**	**0.034**	**0.142**	**0.664**	**0.046**	**0.035**	**5.21**	**0.038**

Table 1: **Results on Simulated and Real-world Datasets.** The Simulated Dataset includes the SoundSpaces dataset, featuring novel views collected in both the Same Environment and Novel Environment. The Real-world Dataset includes the Replay dataset, which contains novel views in the same environment, and the EnvSound dataset, which we collected ourselves across various environments. The Replay dataset was collected using a professional setup, whereas the EnvSound dataset was captured using a phone recorder. We consider the metric of STFT magnitude (Mag), left/right ratio error (LRE), and RT60 error (RTE). We evaluate the novel environment for the SoundSpaces dataset for it has a subset rendered on novel scenes. Lower is better for all baselines, for which we use the input audio, TF estimator, DSP, and the ViGAS model.

5.2 Model Evaluations

We evaluated our model's performance on three datasets. For the simulated SoundSpaces dataset, we tested our model on uncaptured views from both "Same Environment" and "Novel Environment". We randomly activate only one speaker and select two viewpoints as the source and target during training. For the real-world Replay dataset, which provides only a single environment, we used uncaptured views from the same environment. The EnvSound dataset consists of diverse everyday scenarios which we use to evaluate our model's robustness in challenging environments which include background music and natural ambient noise. All models were trained on RTX3090 GPUs. Table 1 shows the performance of the models on each of these datasets.

We compared our model against four baselines: 1) audio-only input , where the synthesized audio at all viewpoints is identical to the input source audio, 2) TF estimator [68], 3) digital signal processing (DSP) [7], and 4) ViGAS model [6] without the depth map. The DSP baseline predicts the output audio in two stages: first, converting the binaural audio from the source location to the speaker's mono audio using the inverse head-related transfer function (HRTF), and second, applying the target microphone pose and HRTF to process the speaker's audio and obtain the final output. For the TF Estimator, we use a Wiener filter to estimate and store transfer functions, indexed by location or pose, and retrieve the nearest match at test time. We supplied the ground-truth coordinates from the SoundSpaces dataset for the DSP baseline. As can be seen in Table 1, our method demonstrate consistently better performance over all baseline methods and across synthetic and real data. As our custom EnvSound data is captured in uncalibrated in-the-wild environments, we are unable to evaluate traditional methods that require speaker coordinates, such as TF Estimator and DSP, on this dataset. Our enhancements across all metrics compared to existing methods indicate that our method is robust diverse scenarios and everyday noisy environments. We provide additional results and video evaluations in the Supplementary Material.

We used three metrics to measure the deviation of the predicted audio from the ground truth,

- STFT Magnitude Error (Mag) compares the magnitude of the Short-Time Fourier Transform (STFT) between the predicted and ground truth audio signals. A lower Mag error suggests better preservation of the audio's frequency components. Early reflections, occurring within the first 50 ms, demand high temporal resolution for accurate analysis. By using a hop length of 160 samples at a 16,000 Hz sampling rate, we achieve a 10 ms update interval, enabling precise capture of their timing and characteristics.
- RT60 Error (RTE) measures the difference in reverberation time (RT60) between the predicted and ground truth audio signals. A lower RTE indicates that the model accurately captures the reverberation properties of the environment.
- Left-Right Ratio Error (LRE) calculates the difference in energy ratio between the left and right channels of the predicted and ground truth audio signals. A lower LRE suggests that the model accurately reproduces the spatial balance between channels, crucial for creating a realistic binaural audio experience.

These metrics collectively evaluate the spectral content, reverberation characteristics, and spatial balance of the predicted audio, assessing how well the model captures the acoustic properties of the environment and reproduces the desired binaural audio output.

Overall, the proposed MNAF framework effectively synthesizes audio at novel viewpoints as also demonstrated by the objective metrics in reported in Table 1. Additionally, we also showcase in Figure 6 a comparison between the predicted novel view audio waveform by our method, by the best available alternate method ViGAS, and the groundtruth audio waveform. This demonstrates that out method outperforms prior approaches and achieves audio synthesis that matches closely with the groundtruth. Our subjective user evaluations, which also demonstrate significantly improved immersive audio perception, are discussed in Section 7 and Section 8.

6 Analysis

In this section, we analyze our framework and conduct ablation studies to evaluate the contribution of each network component towards its overall effectiveness. Specifically, we study the contribution of the conformer (used for visual-acoustic fusion) by replacing it with a naive concatenation of visual and coordinate features, and adaptive convolutions (used for acoustic synthesis) by substituting them with standard convolutions. Our tests validate the choice of our network architecture's visible improvements in performance.

6.1 Analysis on Synthetic Data

Our method outperforms all baseline approaches, including the traditional approaches that consider explicit camera and speaker location coordinates, on the synthetic SoundSpaces dataset. To assess the contributions of individual network components, we conducted ablation tests on the choice of conformer and adaptive convolutions, and the results are shown in Table 2.

Specifically, we test on two different scenarios: acoustic synthesis at a novel viewpoint in the same environment and acoustic synthesis in an unseen novel environment. We observe that the adaptive convolutions significantly improve STFT magnitude and left-right ratio by capturing the long-term dependencies at higher adaptive resolutions. The conformer, on the other hand, shows visible improvements across all metrics, especially in RTE that measures reverberation. A similar trend is observed even in novel environments, where our method shows significant improvements over prior works. The adaptive convolutions demonstrate improvements over all metrics and shows our method's adaptability to different unseen novel environments. The conformer improves on the STFT magnitude and LRE but show modest improvements for RTE. This is because the encoded reverberation patterns of one environment might not scale well to novel unseen environments.

Fig. 6: **Perceptual Study Results.** We present comparisons of predicted novel viewpoint audio for our model to the ViGAS model on both channels. The speaker and the position of the novel viewpoint at which audio is predicted are indicated in the figures. The waveforms predicted by our model are closer to the ground truth. More results, including waveform plots and videos, are shown in the Supplementary Material.

Models	Same Environment			Novel Environment		
	Mag	LRE	RTE	Mag	LRE	RTE
Our Model	**0.134**	**1.151**	**0.034**	**0.186**	**1.125**	**0.034**
Our Model w/o Conformer	0.139	1.174	0.037	0.191	1.21	0.034
Our Model w/o Adaptive Convolutions	0.150	1.260	0.037	0.211	1.241	0.036

Table 2: **Ablation on Synthetic Dataset.** Ablation tests and analysis of our method on synthetic data in rendering sound in familiar and unseen novel environments.

6.2 Analysis on Real-World Data

For testing the performance of our method on real data, we evaluate it on the Replay dataset and show that our model outperforms the baselines. For the Replay dataset, adaptive convolutions contribute more significantly to improving the left-right energy ratio compared to simulated datasets. Real-world audio features a wider range of acoustic phenomena such as reflections and subtle diffractions that are not often not well modeled in simulated datasets. These complexities create diverse and intricate spatial audio cues that adaptive convolutions can effectively model. As Replay dataset consists of clips within the same environment, we see an improvement on RTE contributed by the conformer (Table 3) despite the lack of ground truth depth values. Our method requires at least two viewpoints and increasing the number of viewpoints empirically results in more accurate outcomes. Results on real-world datasets validate the effectiveness of our proposed approach in capturing room acoustics and synthesizing spatial audio in everyday environments. Additional results and analysis can be found in the Supplementary Material.

Models	Mag	LRE	RTE
Our Model	**0.142**	**0.664**	**0.046**
Our Model w/o Conformer	0.143	0.664	0.0475
Our Model w/o Adaptive Convolutions	0.147	0.804	0.049

Table 3: **Ablation on Realistic Dataset.** Ablation tests and analysis of our method on real-world dataset (Replay).

6.3 Preliminary Validation Towards Personalization

Akin to today's mobile phones, augmented reality headsets of the future will be personalized and that requires personalizing spatial audio rendering by incorporating the HRTF of the individual users. To explore this possibility, we evaluated the integration of an additional HRTF module to our current model for end-to-end learning of binaural audio. To this end, an HRTF module that takes the speaker's absolute position and pose relative to the target view is added before the visual-acoustic fused signals are passed to the acoustic synthesis block. This enables predicting binaural signals by conditioning the audio synthesis block with re-encoded binaural positional cues. Our current HRTF module employed the HRTF filter of the KEMAR dummy head [14], incorporating the speaker's absolute position and pose relative to the target view. User-specific HRTFs, which can vary significantly between individuals, can be derived through methods such as 3D head scanning [1, 56] or using personal anthropometric measurements [71, 74] with a large dataset [66]. However, these methods raise privacy concerns due to the exposure of personal identity. An end-to-end approach to predicting HRTF filters from head movements and binaural audio presents a potential solution to mitigate such privacy risks [19, 30], and is a key direction for our future research. The results shown in Table 4 and Table 5 demonstrate that the HRTF module improves left-right ratio as expected, and also improves phase and delay, indicating its ability to increase spatial quality of predicted audios. Furthermore, the HRTF module also adapts well to novel environments, demonstrating its capability to capture spatial sound fields effectively even in previously unseen settings, as shown in Table 5. Additional discussion on HRTF module can be found in the Supplementary Material.

Models	Mag ↓	RTE ↓	LRE ↓	Phase ↓	Delay ↓
Our Model	0.134	0.034	**1.134**	1.629	2.683
Our Model w/ HRTF Module	**0.134**	**0.034**	1.15	**1.420**	**1.815**

Table 4: **HRTF module in SoundSpaces - Same Environment**

Models	Mag ↓	RTE ↓	LRE ↓	Phase ↓	Delay ↓
Our Model	0.186	0.034	1.112	1.500	2.573
Our Model w/ HRTF Module	**0.185**	**0.034**	**1.053**	**1.46**	**1.815**

Table 5: **HRTF module in SoundSpaces - Novel Environment**

7 Perceptual Quality in VR

As shown in Figure 6, we developed a study to assess the audio rendering quality of unseen viewpoints in live recordings rendered in a VR environment. The aim is to determine how well our model generates realistic audio interpolation in common, everyday settings.

Participants. We recruited 12 subjects (ages 18 - 27, 6 females, 6 males) to join the study. All have normal or corrected-to-normal vision and no history of auditory deficiency. Among the participants, 5 had no prior experience with VR, 5 used VR equipment 5 times or less, and 2 are familiar with and have regular access to VR devices. None were aware of the hypothesis, the research, or the number of task difficulty levels. The study protocol was approved by the Institutional Review Board (IRB) at the host institution, and all subjects gave informed consent before the study.

Setup. In this study, we use a total of 11 scenes, including 7 from the EnvSound dataset as detailed in Section 5.1, along with 4 scenes from the publicly available AVSpeech dataset [11], which features clean speech video clips without background noise. All videos were converted into 180-degree stereoscopic 3D videos with accompanying audio and played in a Meta Quest 3 headset, where the experiment took place.

Stimulus. The stimuli comprise 11 novel view videos featuring everyday conversations in various settings, including noisy, quiet, spacious, and compact rooms. While the visual frames remain unchanged, the audio for each scene is generated using one of three methods: captured audio (reference), audios interpolated by our model, and baseline method [6]. These audio tracks are then synchronized with the visual frames, ensuring consistent visuals and the resulting videos are presented to the participants.

Task. In the VR headset, participants first listened to 11 reference audios derived from novel view frame videos. They were asked to review and memorize these reference sounds before starting the experiment. In each trial, a stimulus generated by one of three rendering methods was presented for 15 to 30 seconds. Participants were instructed to focus on the speaker to ensure consistent sound directivity and used the keyboard to indicate whether the stimulus matched the reference audio. The experiment comprised 33 trials in total, with 11 trials per rendering method, presented in a randomized order.

case \ user	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12
Reference Video	7	6	9	5	4	8	8	4	9	8	11	4
Ours	9	7	8	8	5	6	10	5	9	9	10	4
Baseline	8	9	2	5	0	2	0	2	5	3	8	1

Table 6: **User Study Results.** The table shows the number of trials (out of 12) where participants did not notice any difference compared to the reference. Notably, some participants reported audible difference even in the reference video condition, indicating individual differences in judgment criteria. A significant difference in perceived audio quality was observed between reference vs. baseline and our model vs. baseline conditions ($p = 0.006$). In contrast, the difference between our model and reference was not significant ($p = 0.189$).

Result. A one-way within-subjects ANOVA revealed a significant difference in perceived sound quality between the acoustic synthesis methods ($F_{(2,22)} = 7.39$, $p = 0.002$). Post-hoc paired t-tests with Bonferroni correction showed that the baseline method performed significantly worse than both the reference sound and our model ($p = 0.006$). Importantly, there was no significant difference between the sound of our model and the reference sound ($p = 0.189$). These results suggest that our method achieves sound quality similar to the reference, outperforming the baseline.

8 Evaluation of Immersive Audio

We aimed to evaluate our model's ability to transform existing audio into novel virtual environments, comparing it to prior methods. To this end, we conducted two studies in both VR and AR. Participants first watched a concert video and memorized the reference audio. Then, they experienced the transformed audio generated by our model and prior work [6], which was adapted to the VR/AR environment. User experience questionnaires were provided to assess the participants' immersion and spatial audio sensation of the audio events synthesized between models.

Participants We recruited 6 participants (ages 18 - 27, 1 female) to join the study. All have normal or corrected-to-normal vision and no history of auditory deficiency. Among the participants, 2 had no prior experience with VR, 3 used VR equipment 5 times or less, and 1 is familiar with and has regular access to VR devices. None were aware of the hypothesis, the research, or the number of task difficulty levels. The study protocol was approved by the Institutional Review Board (IRB) at the host institution and all subjects gave informed consent before the study.

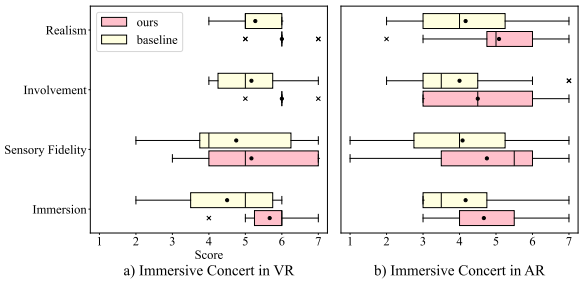
a) User Experience in VR b) User Experience in AR

Fig. 7: **Live Views.** Participants in a) VR, and b) AR environments.

Experimental Setup. We curated live concert clips for reference audio. In the VR study, we created two virtual environments in Unity. For the AR study, participants focused on virtual speakers placed in an office space. The VR study was conducted using the Meta Quest 2 headset, while the AR study involved participants wearing the Apple Vision Pro.

8.1 Application in VR

We adopted two sizes of indoor scenes in the VR application study: an apartment (large) and a bedroom (small). Participants watched the reference audio video before the experiment. Then, they were instructed to walk through two scenes with immersive spatial audio in VR. The VR experience session was repeated twice, with one using the audio event generated by the baseline model [6] and one using our proposed model. After completing the auditory immersion task, participants were asked to fill in 2 questionnaires to complete a post-study survey on evaluating the user experience between 2 sessions (our model vs baseline). The survey consisted of seven questions sourced from standard XR questionnaires: PQ and IPQ. The entire study, including instructions and the survey, took approximately 5 minutes per participant. As shown in Figure 8, post-hoc paired t-tests at the 10% significance level shows that our method outperforms the baseline in realism and immersion for VR ($p = 0.054, 0.002$ respectively).



a) Immersive Concert in VR b) Immersive Concert in AR

Fig. 8: **Subjective evaluation of Immersive Audio in VR/AR.** Our post-study survey includes questions (detailed in Supp.) from the PQ and IPQ questionnaires. The Likert-scale responses are combined to evaluate both conditions across four dimensions: Immersion, Sensory Fidelity, Involvement, and Realism, with higher scores reflecting better performance. The boxes represent the data range between the first and third quartiles (Q1-Q3), with dots showing the mean and lines indicating the median. Error bars represent the farthest data points within 1.5× the interquartile range (IQR) from the boxes.

8.2 Application in AR

We used a conference table as the physical proxy and mapped a designated area in AR to place the virtual speaker. The study was implemented using Apple Vision Pro. The reference audio along with the captured image of the office space, was fed into both our model and the baseline (consistent with Section 8.1). Participants were seated in a chair facing the virtual speakers placed on the conference table. After hearing the reference audio, they were asked to listen to two audio clips played by Apple Vision Pro: the audio generated by our model and the baseline, presented in random order. The same post-study survey was utilized, as outlined in Section 8.1. The entire study, including study instructions and the post-study survey, took approximately 5 minutes per participant. Post-hoc paired t-tests at the 10% significance level, as shown in Figure 8, demonstrate that our method performs better in Involvement, Sensory Fidelity and Immersion for the AR setting ($p = 0.027, 0.056, 0.007$).

8.3 Results

For the VR Concert application, we demonstrate consistent enhancement in the feeling of presence over the baseline, particularly in immersion and realism. In VR, Immersion improves by 26% in mean and reduces variance by 51.4%, indicating a more stable experience. Realism sees a 13.4% mean improvement and a 34.5% variance reduction, ensuring greater reliability. In AR, Immersion improves by 12% in mean and 19.7% in variance, while Realism shows a 14.1% mean boost and a notable 43.7% variance reduction.

9 Discussion and Future Work

We introduced a multimodal neural acoustic field framework that synthesizes spatial sound and enhances immersive auditory experiences in virtual and augmented reality environments by mapping geometric and visual features to audio characteristics. Using a hybrid transformer-convolutional neural network and an adaptive convolution-based acoustic synthesis module, our model can capture reverberation and generate spatial sound from sparse signals in unseen novel environments. Our approach improves spatial audio quality and realism, validated through analysis on synthetic and real-world data as well as subjective user studies, particularly benefiting augmented and virtual reality applications. Additional user evaluations with a more diverse participant pool will be valuable to further assess the generalizability of our approach. Our current implementation does not handle 360-degree video inputs and struggles in scenarios where a meaningful geometric or material context cannot be derived, such as images featuring plain white walls or blank spaces. Extending the current neural networks to explicitly model and learn the material properties of objects for more accurate synthesis of acoustic effects is a promising future direction. Additionally, we plan on expanding experimental conditions to include dynamic scenes with multiple moving sources and extremely noisy scenarios. To show real-world adaptability, we would also test the method on various AR/VR platforms and monitor its power consumption. As such, this paper paves the road toward multimodal neural rendering for mixed reality of the future, and we are excited that our work will inspire further investigation into immersive and personalized audio-visual experiences.

References

[1] Listen with personalized spatial audio for airpods and beats. 7

[2] E. H. Adelson. On seeing stuff: the perception of materials by humans and machines. In *IS&T/SPIE Electronic Imaging*, 2001. 3

[3] J. Bradley. Review of objective room acoustics measures and future needs. *Applied Acoustics*, 72(10):713–720, 2011. doi: 10.1016/j.apacoust.2011.04.004 2

[4] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 5

[5] C. Chen, R. Gao, P. Calamia, and K. Grauman. Visual acoustic matching, 2022. 2, 3

[6] C. Chen, A. Richard, R. Shapovalov, V. K. Ithapu, N. Neverova, K. Grauman, and A. Vedaldi. Novel-view acoustic synthesis. In *CVPR*, 2023. 2, 3, 4, 5, 6, 8

[7] C. Cheng and G. Wakefield. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. *AES: Journal of the Audio Engineering Society*, 49:231–249, 04 2001. 6

[8] N. Deng, Z. He, J. Ye, B. Duinkharjav, P. Chakravarthula, X. Yang, and Q. Sun. Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3854–3864, 2022. 2

[9] C. Donahue, J. McAuley, and M. Puckette. Adversarial audio synthesis, 2019. 2

[10] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts. Gansynth: Adversarial neural audio synthesis, 2019. 2

[11] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 8

[12] D. H. Foster, S. M. Nascimento, K. Amano, L. Arend, K. J. Linnell, J. L. Nieves, S. Plet, and J. S. Foster. Parallel detection of violations of color constancy. *Proceedings of the National Academy of Sciences*, 98(14):8151–8156, 2001. 3

[13] R. Gao and K. Grauman. 2.5d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5

[14] B. Gardner. Hrtf measurements of a kemar dummy-head microphone. 1994. 7

[15] M. B. Gardner and R. S. Gardner. Problem of localization in the median plane. *The Journal of the Acoustical Society of America*, 51(1A_Supplement):149–149, 1972. 1

[16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang. Conformer: Convolution-augmented transformer for speech recognition, 2020. 3

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. 3

[18] A. Jain, M. Tancik, and P. Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis, 2021. 2

[19] V. Jayaram, I. Kemelmacher-Shlizerman, and S. M. Seitz. Hrtf estimation in the wild, 2023. 7

[20] Y. Jiang, J. Tu, Y. Liu, X. Gao, X. Long, W. Wang, and Y. Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces, 2023. 2

[21] Y. Jiang, C. Yu, T. Xie, X. Li, Y. Feng, H. Wang, M. Li, H. Lau, F. Gao, Y. Yang, et al. Vr-gs: a physical dynamics-aware interactive gaussian splatting system in virtual reality. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–1, 2024. 2

[22] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 2

[23] I. Khalfaoui-Hassani, T. Pellegrini, and T. Masquelier. Dilated convolution with learnable spacings, 2023. 4

[24] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2021. 2

[25] A. Krokstad, S. Strom, and S. Sørsdal. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118–125, 1968. 1, 2

[26] J. Kulhánek, E. Derner, T. Sattler, and R. Babuška. Viewformer: Nerf-free neural rendering from few images using transformers, 2022. 2

[27] H. Kuttruff. *Room Acoustics*. CRC Press, 6 ed., 2016. doi: 10.1201/9781315372150 2

[28] G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding, 2017. 3

[29] Z. Li, S. Niklaus, N. Snavely, and O. Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes, 2021. 2

[30] Z. Li, B. Zhao, and Y. Yuan. Cyclic learning for binaural audio generation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26669–26678, June 2024. 7

[31] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis, 2023. 3

[32] L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, and C. Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control, 2022. 2

[33] S. Liu and J. Liu. Outdoor sound propagation based on adaptive fdtd-pe. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 859–867. IEEE, 2020. 2

[34] M. Long. *Architectural acoustics*. Academic press, 2014. 3

[35] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan. Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, 35:3165–3177, 2022. 2

[36] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio. Samplernn: An unconditional end-to-end neural audio generation model, 2017. 2

[37] L. Mei, S. Wang, Y. Cheng, R. Liu, Z. Yin, W. Jiang, S. Wang, and W. Gong. Esp-pct: Enhanced vr semantic performance through efficient compression of temporal and spatial redundancies in point cloud transformers. In K. Larson, ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 1182–1190. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track. doi: 10.24963/ijcai.2024/131 2

[38] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[39] M. Morimoto and Y. Ando. On the simulation of sound localization. *Journal of the Acoustical Society of Japan (e)*, 1(3):167–174, 1980. 1

[40] I. Motoyoshi, S. Nishida, L. Sharan, and E. H. Adelson. Image statistics and the perception of surface qualities. *Nature*, 447(7141):206–209, 2007. 3

[41] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. M. Sajjadi, A. Geiger, and N. Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs, 2021. 2

[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. IEEE, 2015. 5

[43] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Nerfies: Deformable neural radiance fields, 2021. 2

[44] S. Pascual, G. Bhattacharya, C. Yeh, J. Pons, and J. Serrà. Full-band general audio synthesis with score-based diffusion, 2022. 2

[45] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes, 2020. 2

[46] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister. Langsplat: 3d language gaussian splatting, 2023. 2

[47] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction, 2021. 2

[48] A. Richard, P. S. Dodds, and V. K. Ithapu. Deep impulse responses: Estimating and parameterizing filters with deep networks. arXiv preprint arXiv:2202.03416, 2022. 2

[49] A. Rungta, S. Rust, N. Morales, R. Klatzky, M. Lin, and D. Manocha. Psychoacoustic characterization of propagation effects in virtual environments. *ACM Transactions on Applied Perception (TAP)*, 13(4):1–18, 2016. 1

[50] C. Schissler, P. Stirling, and R. Mehra. Efficient construction of the spatial room impulse response. In *2017 IEEE Virtual Reality (VR)*, pp. 122–130. IEEE, 2017. 2

[51] D. Schwarz. Concatenative sound synthesis: The early years. *Journal of New Music Research*, pp. 3–22, 2006. 2

[52] R. Shapovalov, Y. Kleiman, I. Rocco, D. Novotny, A. Vedaldi, C. Chen, F. Kokkinos, B. Graham, and N. Neverova. Replay: Multi-modal multi-view acted videos for casual holography. In *ICCV*, 2023. 5

[53] L. Sharan, Y. Li, I. Motoyoshi, S. Nishida, and E. H. Adelson. Image statistics for surface reflectance perception. *Journal of the Optical Society of America A*, 25(4):846–865, 2008. 3

[54] S. Siltanen, T. Lokki, S. Kiminki, and L. Savioja. The room acoustic rendering equation. *The Journal of the Acoustical Society of America*, 122(3):1624–1635, 09 2007. doi: 10.1121/1.2766781 2

[55] V. Sitzmann, M. Zollhoefer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. 2

[56] R. Sridhar and E. Choueiri. A method for efficiently calculating head-related transfer functions directly from head scan point clouds. 10 2017. 7

[57] C. J. Steinmetz, J. Pons, S. Pascual, and J. Serrà. Automatic multitrack mixing with a differentiable mixing console of neural audio effects. In *ICASSP*, 2021. 5

[58] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5

[59] K. Su, M. Chen, and E. Shlizerman. Inras: Implicit neural representation for audio scenes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds., *Advances in Neural Information Processing Systems*, vol. 35, pp. 8144–8158. Curran Associates, Inc., 2022. 2

[60] S.-Y. Su, F. Yu, M. Zollhoefer, and H. Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose, 2021. 2

[61] Z. Tang, H.-Y. Meng, and D. Manocha. Learning acoustic scattering fields for dynamic interactive sound propagation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 835–844. IEEE, 2021. 1, 2

[62] L. L. Thompson. A review of finite-element methods for time-harmonic acoustics. *The Journal of the Acoustical Society of America*, 119(3):1315–1330, Mar. 2006. doi: 10.1121/1.2164987 2

[63] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Lassner, and C. Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video, 2021. 2

[64] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio, 2016. 2

[65] M. Vorländer. Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm. *The Journal of the Acoustical Society of America*, 86(1):172–178, July 1989. doi: 10.1121/1.398336 2

[66] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato. Dataset of head-related transfer functions measured with a circular loudspeaker array. *Acoustical Science and Technology*, 35:159–165, 05 2014. doi: 10.1250/ast.35.159 7

[67] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123, 1993. 1

[68] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964. 6

[69] R. Yamamoto, E. Song, and J.-M. Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203, 2020. doi: 10.1109/ICASSP40776.2020.9053795 5

[70] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelnerf: Neural radiance fields from one or few images, 2021. 2

[71] M. Zhao, Z. Sheng, and Y. Fang. Magnitude modeling of personalized hrtf based on ear images and anthropometric measurements. *Applied Sciences*, 12(16), 2022. doi: 10.3390/app12168155 7

[72] C. Zhou, M. Horgan, V. Kumar, C. Vasco, and D. Darcy. Voice Conversion with Conditional SampleRNN. In *Proc. Interspeech 2018*, pp. 1973–1977, 2018. doi: 10.21437/Interspeech.2018-1121 2

[73] G. Zhu, Y. Wen, M.-A. Carbonneau, and Z. Duan. Edmsound: Spectrogram based diffusion models for efficient and high-quality audio synthesis, 2023. 2

[74] D. Zotkin, J. Hwang, R. Duraiswaini, and L. Davis. Hrtf personalization using anthropometric measurements. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, pp. 157–160, 2003. doi: 10.1109/ASPAA.2003.1285855 7