

# Materialistic RIR: Material Conditioned Realistic RIR Generation

## Supplementary Material

### 7. Supplementary Material

In this supplementary material, we provide further analyses and details about our approach, and we will release code and data upon acceptance.

- Supplementary video (with audio) for qualitative results, Sec. 7.1
- Details of the proposed evaluation metrics for measuring our predictions’ ability to capture the material-dependence of scene acoustics (Sec. 7.2) as mentioned in L399 and L414 in Sec. 4.4 in main
- Computation cost analysis of our model and the baselines (Sec. 7.3) as stated in L537 in Sec. 5.3 in main
- Additional Analysis of our model for fine-grained changes to the material mask (Sec. 7.4)
- Model architecture (Sec. 7.5) as mentioned in L381 in Sec. 4 in main
- Limitations and Discussion (Sec. 7.6)
- Real-World Experiments (Sec. 7.7)
- Details about Perceptual Evaluation of Our Model using Humans (Sec. 7.8), as mentioned in Sec. 5.2 in main
- Evaluation Setup (Sec. 7.9) as mentioned in Sec. 5 in main

#### 7.1. Supplementary Video

We provide a supplementary video with audio to qualitatively show our model outputs. Specifically, we show qualitative examples, which are speech samples from Librispeech[54] dataset, convolved with predicted RIRs from our model, compared against the same when convolved with the outputs of our best baselines. The results demonstrate the ability of our approach to model the fine-grained changes in *both* the material and spatial layout in the scene.

#### 7.2. Metrics for Evaluating Material Specificity

In this section, we provide additional details for our proposed evaluation metrics for capturing our model’s material awareness, MatC and MatD.

##### 7.2.1. Material Classification Accuracy (MatC)

The MatC metric measures the effectiveness of a method’s ability to model the material impact on RIRs in single-material setups. In particular, this metric leverages a network that is trained to predict the material type from a given RIR, where all objects in the scene are assigned a single material. To compute this metric, we freeze the above-mentioned material classifier, run inference with it and compute its accuracy vis-a-vis predicting the expected material type as its output.

We pre-train the material classification network using the SSv2 [10] simulator and the MP3D [6] dataset. We use the same splits as M-CAPA [62] with 76, 3 and 5 scenes in train, val and test, respectively. Additionally, the total ground-truth RIR count is 167K for train, 6.6K for val, and 11K for test. To generate these RIRs, we randomly sample a single material class from the AcoW [62] dataset, assign it to all objects in a scene, and finally capture the echo responses from different locations and viewpoints in the scene. Using this data, we train a ResNet18 [26] architecture to take 2-channel binaural RIRs, represented as spectrograms, and classify it into one of the 11 material classes from AcoW. During training, we apply Gaussian noise to the RIRs to ensure the classifier is robust to microphone noise [32, 70]. We train the classifier for 50 epochs with the Cross-Entropy loss [50] and a learning rate of 0.01. This classifier is able to achieve 96.7% accuracy on the test set.

##### 7.2.2. Material Distribution Accuracy

This metric is used to measure the accuracy with which a model can capture in its predicted RIR the acoustic signature of a distribution of different materials in the scene. Here, we use a network that is pretrained to predict the material distribution in the scene given an RIR, where (unlike the case with MatC) the various objects in the scene can have different materials.

For this metric, we use the same training, validation, and test splits as described in Sec 5. To pretrain this network, we first extract the material class distribution obtained from the material segmentation mask  $M$  as a percentage of area covered for each material class. Next, we group these distributions into 36 distinct clusters using K-Means [66]. This helps us group scenes that have similar material configurations together. We then train a ResNet18 network [26] such that for a given binaural RIRs, it will predict the cluster index that corresponds to the material distribution associated with that input RIR. This cluster classification network is trained for 10 epochs with the Cross-Entropy loss [50] and a learning rate of  $1e^{-3}$ . This network achieves 77% Top-5 accuracy on the test split. For computing this metric, we freeze the network, run inference with it and measure its top-5 accuracy vis-a-vis predicting the expected material distribution cluster as its output.

#### 7.3. Computational Costs

Table 3 shows a comparison of our model and the baselines on the basis of the number of trainable parameters, GFLOPs and inference time. Note that when counting our model

Method	Params (M)	GFLOPs	Inf. Time (ms)
Image2Reverb [67]	57.6	276.91	198.44
FAST-RIR[58]++	132.68	57.84	121.76
M-CAPA	5.84	11.24	76.61
JM-CNN	14.99	10.71	187.86
JM-Enc	6.22	6.21	227.81
JM-QFormer	7.34	7.29	293.10
MatRIR (Ours)	13.28	14.11	270.56

Table 3. Computational cost of our approach (MatRIR) and the baselines.

parameters, we exclude the modules like DinoV2 [53] and Midas [4], as they are kept frozen during training.

## 7.4. Additional Model Analyses

In this section, we conduct further analysis of our model’s ability to capture fine-grained variations in the spatial layout as well as material configurations of the input scenes. Furthermore, we study the impact of the type of the dominant material in the input material mask on the model performance.

### 7.4.1. Fine-Grained Environmental Changes

Our approach to material-conditioned RIR generation explicitly disentangles the spatial and material components of the room impulse response. This disjoint modeling enables fine-grained and interpretable control, allowing us to modify material properties of the scene while preserving the underlying geometric layout of the scene. Our method uses this modular approach to generate accurate and meaningful variations in environmental acoustics, conditioned on unique material properties.

To this end, we show qualitative examples in Fig. 6, where we demonstrate the fine-grained control over material variations provided by our model. For each example, we use two distinct material configurations while keeping the spatial layout of the scene fixed. The RIR predictions for these examples highlight that our model consistently captures the spatial aspects of the ground-truth RIR, while also successfully incorporating the material properties of the target materials in  $M$ . In particular, our approach reliably modifies the spatial RIR  $\hat{A}_S$ , modulating reverberation patterns to accurately reflect the material configuration of the environment, producing RIRs closer to the ground-truth every time. In contrast to our approach, M-CAPA grants limited control on environmental material changes, generating nearly identical RIRs across different spatial and material configurations. For instance, in the first example, when the walls change from concrete to brick, our model modulates the spatial RIR  $\hat{A}_S$  to accurately predict the material-conditioned RIRs  $\hat{A}_M$ , whereas M-CAPA predictions remain relatively unchanged across these two material settings.

### 7.4.2. Material Composition Analysis

Here, we evaluate how difficult each material from the AcoW dataset is for our model to capture in its predicted RIRs. To this end, we assign all objects in an input to a single material and plot in Fig. 7a how the model performance changes in relation to the assigned material. Specifically, the figure shows the value of the MatC metric—material classification (cf. Sec. 7.2.1) accuracy—for every material in the AcoW dataset. Notably, the model can accurately capture the acoustics for most materials; however, *steel* seems to be a difficult class for our model to acoustically model.

We evaluate our model’s effectiveness in capturing a mixture of material distributions in the predicted RIR by using the MatD metric (cf. Sec. 7.2.2). Figure 7b shows the accuracy of this metric with respect to various distribution of materials over the 11 material classes in the dataset. We use the cluster center (see Sec. 7.2.2) as the representative of each group of similar material distributions. We have 36 unique material distributions in total. Our model is able to capture acoustic characteristics of the scene with distribution over materials in cluster 1 and 31 accurately. Analyzing the most prominent materials at these cluster centers, we find that cluster 1 contains large distribution of *sound-proof* material at 44% and *glass* material at 24%, and cluster 31 contains large distribution of *steel* at 43% and *glass* at 16%. Our model can accurately capture the acoustic properties of these mixture of materials in the scene.

On the other hand, we notice that our model performs poorly on cluster 10 which contains mixture of *carpet* at 28% and *concrete* at 29%. We also notice the same performance on cluster 25 with 28% *acoustic tile* material and 27% *grass* material and cluster 34 with 28% *concrete* and 27% *brick*. These mixture of materials in the scene seems difficult for our model to capture acoustically, resulting in a 0% accuracy on the MatD metric for these particular distributions.

## 7.5. Model Architecture

In this section, we provide further details about our model’s architecture and training.

**Depth Predictor** We use a pretrained model from [4] to obtain the depth map input for our model by feeding in our visual representation as an RGB image  $V$  and obtaining a depth map. We normalize this depth map and use it as a grayscale image.

**Spatial Encoder** Our spatial encoder  $\mathcal{E}_S$  is a pretrained, frozen DINOv2-Large [53] encoder that takes visual representations RGB ( $V$ ) and depth map ( $D$ ) and produces a feature maps  $e_v, e_d \in \mathbb{R}^{256 \times 1024}$  extracted from the 18th layer, where each feature map has 256 tokens and is 1024-dimensional.

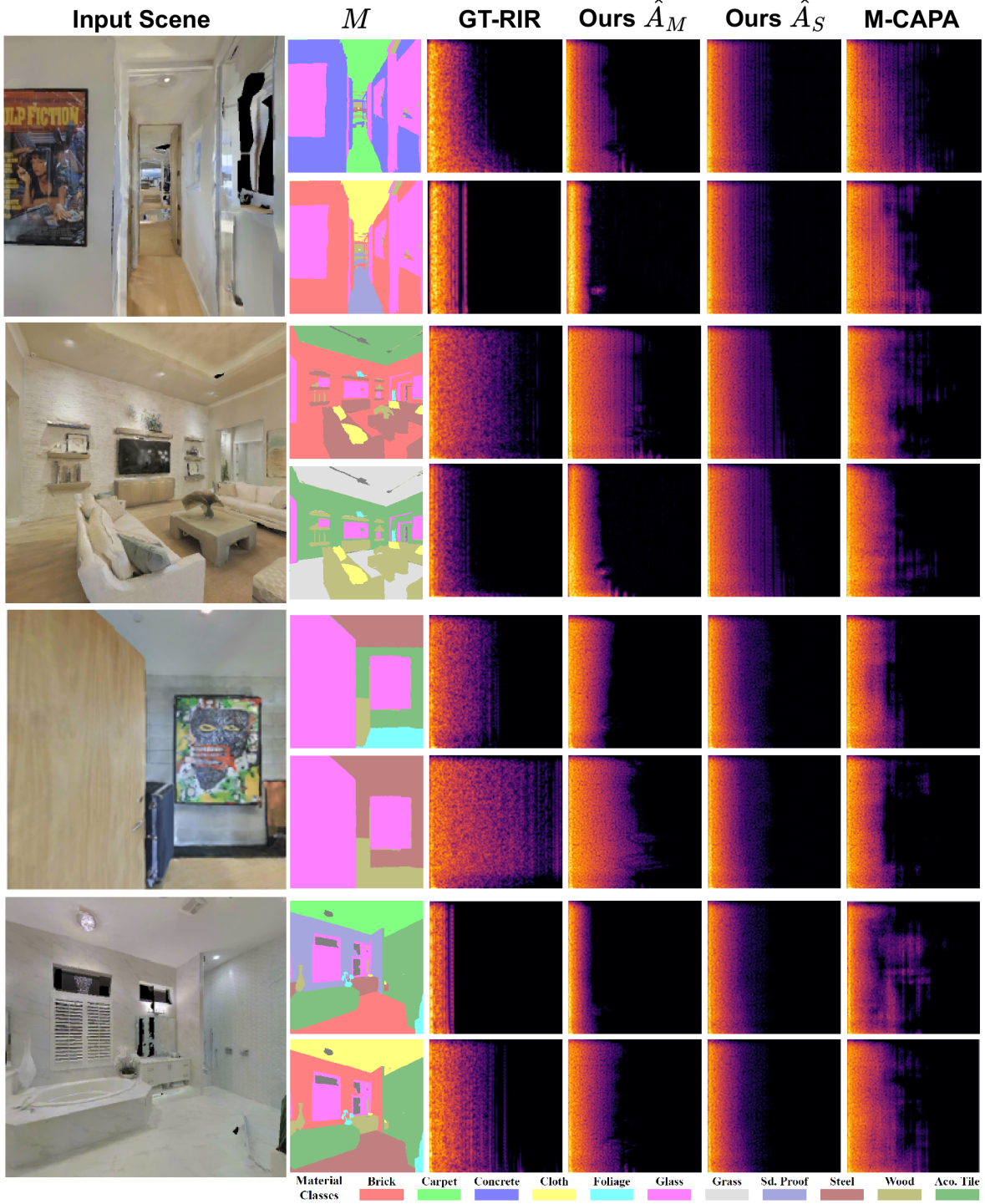
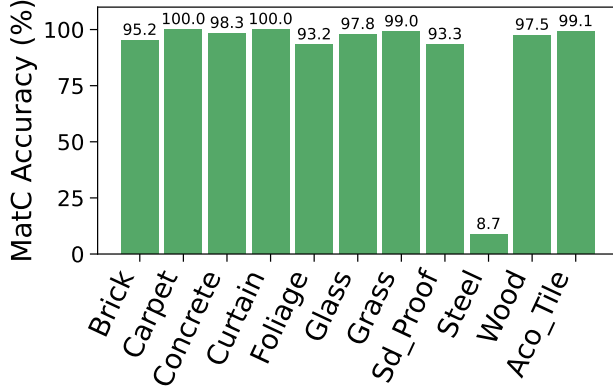


Figure 6. Sample outputs of our model in 4 different scenes. Our model is able to accurately capture acoustic changes conditioned on the given materials configuration. For brevity, we only show one channel of the binaural RIR  $\hat{A}_M$ .

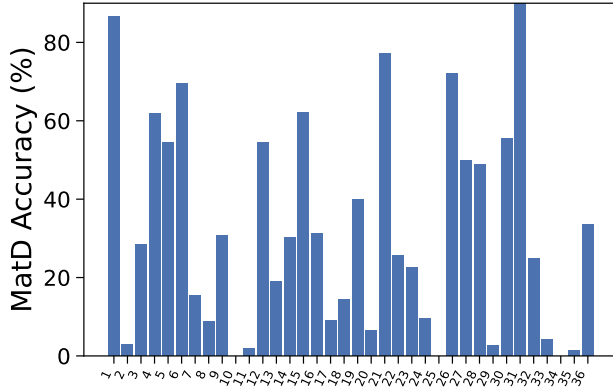
**Material Encoder** Our material encoder  $\mathcal{E}_M$  is a pre-trained, frozen DINOv2-Large [53] encoder that takes material mapping  $M$  as input and produces a feature map

$e_m \in \mathbb{R}^{256 \times 1024}$  from the 18th layer, where each of the 256 tokens is represented by 1024 channels.





(a) MatC accuracy vs. material class



(b) MatD accuracy vs. material distribution

Figure 7. Evaluation of acoustic modeling with respect to the type and distribution of material present in the target material mask.

**Spatial RIR Decoder** To learn spatial acoustic features, we follow Blip-2 [38] architecture for our Spatial RIR Decoder  $\mathcal{R}_S$ . We implement a 4-layer transformer decoder with 256 dimensionality, feedforward layers of 512 dimensionality, and a dropout rate of 0.1. To identify each input modality in this decoder, we follow [49] by using modality-specific embeddings  $s_v$  for vision modality  $V$  and  $s_d$  for depth modality  $D$ . These modalities are concatenated with their respective modality features  $e_v, s_v$  and  $e_d, s_d$ , and passed through a projection layer to bind modality-specific features. The final modality-specific feature tokens are concatenated into a single sequence and projected to decoder dimensionality  $f \in \mathbb{R}^{512 \times 256}$  and used as keys and values. Meanwhile, spatial queries of length 256 are used as the queries for our decoder. The output from this module  $\mathcal{R}_S$  is learned tokens  $g_s \in \mathbb{R}^{512 \times 256}$ .

**Material RIR Encoder** This module,  $\mathcal{R}_M$ , conditions the spatial RIR on material information. We follow a 4-layer transformer encoder [18] with a 256-dimensionality, feedforward of 512-dimensionality, and a dropout rate of 0.1. We first use a convolutional patch extractor that patchifies

each channel of the spatial RIR  $\hat{A}_S$  separately. We create patches of size  $16 \times 16$  for each channel, then concatenate the tokens and use a 2-layer MLP to project such that an average feature is learned for each corresponding patch of the left and right channel spectrograms. Furthermore, this module uses 4 re-weighting tokens [29] to learn cross-modality modulation between material and spatial acoustics. These tokens are projected to the dimensionality of the encoder such that  $R \in \mathbb{R}^{4 \times 256}$ . Features of  $M, \hat{A}_S, R$  are concatenated and fed as input to  $\mathcal{F}_M$ , which performs self-attention across the input space. Next, we extract encoded audio tokens  $g_m$  and re-weighting tokens  $g_r$  from the encoder and feed them into their respective modules.

**Audio Feature Upsampling Network** The upsampling network  $\mathcal{U}_S$  consists of 4 transpose convolution layers. Each layer consists of a transpose convolution with a kernel size of 2 and a stride of 2, followed by 2 convolutional layers and leaky ReLU activations [81]. The input feature to the model is upsampled using [512, 256, 128, 64, 32] channels. Finally, a two-layer MLP follows the maps the output to a 2-channel spectrogram of size  $256 \times 256$ .

**Material-Aware Audio Feature Upsampler** Similar to  $\mathcal{U}_S$ , this module  $\mathcal{U}_M$  upsamples feature map from the material encoder  $g_m$ . Features from  $g_r$  are independently projected to each layer of  $\mathcal{U}_M$  using a linear projection followed by sigmoid [24] to modulate the output in consecutive layers with cross-modal reweighting features. The output of this module is the final estimation of the material-aware binaural spectrogram.

**Cross-Modal Correspondence Network** This network uses dual ResNet18 architectures for encoding material map  $M$  and predicted RIR  $\hat{A}_M$ , followed by a 3-layer MLP that classifies whether these two inputs correspond or not. We pretrain this network on the training data from [62] for 10 epochs with Binary Cross-Entropy Loss, where 75% of the batch, during training, consists of negative samples. We achieve an 81% accuracy on the unseen scene and unseen material split. Once trained, we freeze the weights of this network and feed it with  $M$  and predicted RIR  $\hat{A}_M$  from  $\mathcal{F}_M$  and use the output to provide feedback to the model during training. We use Binary Cross-Entropy as an augmented training objective to the main model  $\mathcal{F}$ .

## 7.6. Limitations and Discussion

We show that our disjoint modeling for RIR prediction quantitatively outperforms strong baselines and SOTA, in both acoustic and material-based metrics. Qualitative results show the effectiveness of our approach in producing spatial and material-aware RIRs. However, as we show in Fig. 5 in main,

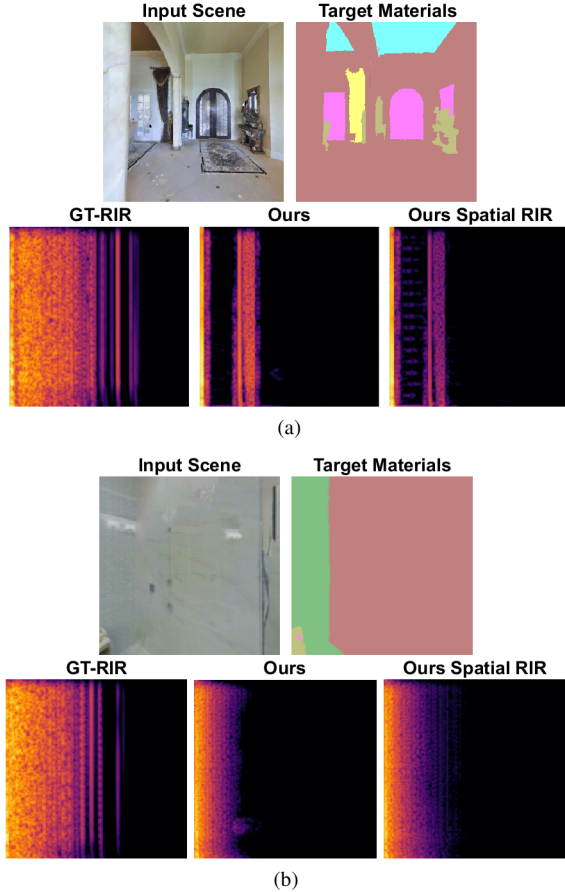


Figure 8. Qualitative examples where our model fails to accurately capture the material-dependence. We observe that in cases where there are big occlusions, and where *steel* is the dominant material in the target material mask, our model struggles to capture any acoustic aspects beyond what its geometry-conditioned initial estimate,  $\hat{A}_S$ , already captures, thereby resulting in a lack of material awareness.

when the scene is not fully visible, or the visual representation is obstructed by large objects (like walls or shelves), our model is not able to accurately capture the material distribution of the scene in the predicted RIR. A possible solution to this type of error is to provide a full panoramic visual and material representation of the scene, such that the model can condition on the full  $360^\circ$  visual input.

Furthermore, from Fig. 7a, we see that our model struggles to model the acoustic properties of *steel*, when everything in the scene is assigned to this material. This observation is supported by the qualitative example in Fig. 8a and 8b where the target material mappings consists of a large amount of *steel* in the scene. Our model struggles to model the material acoustics in this setting.

Generating accurate room acoustics conditioned on the surface materials of the objects within a space is an interesting challenge. While in this work we model the impact of 11 material classes presented in [62], including more material

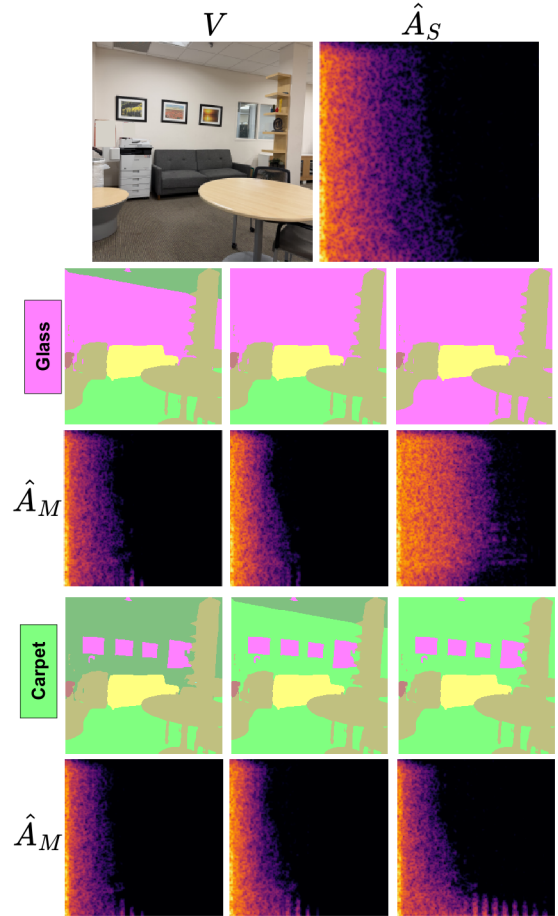


Figure 9. Qualitative examples of our model predictions in a real-world setup. We experiment with different material configurations for the same scene where we increase the presence of one material class (*Glass* top, and *Carpet* bottom) in the target material mask gradually and demonstrate the resulting material-conditioned RIRs,  $\hat{A}_M$  produced by our model.

representations in the dataset would be beneficial in order to capture the large diversity of material distributions in the real world.

## 7.7. Real-World Experiments

To validate the applicability of our approach beyond simulated environments, we qualitatively evaluate our model on real-world scenes. Towards this goal, we collect snapshots of different places in a building and run inference on these images with our model. To do so, we first extract depth maps from the images using Midas [4] and semantic segmentation masks using InternImage [77], and then use these as model inputs in order to obtain the spatially-accurate initial RIR predictions,  $\hat{A}_S$ , and the material-aware final predictions,  $\hat{A}_M$ .

In Fig. 9, we show examples of our model predictions for

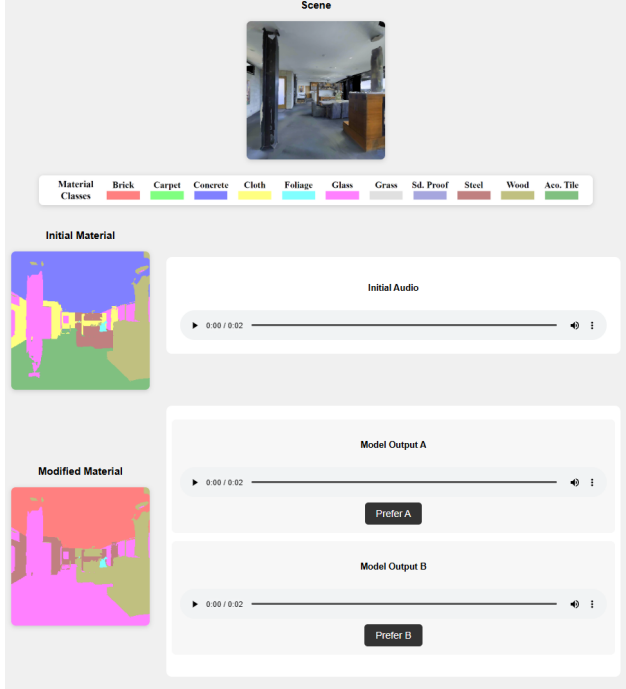


Figure 10. Interface for our user study where participants listen to predicted RIRs from MatRIR and M-CAPA, convolved with clean speech audio, and select which model produces more realistic acoustics.

a randomly-chosen single RGB image, and consequently, a single depth map depicting the spatial layout in the scene, but different target material masks. By modifying the object-to-material assignments in the scene, we demonstrate how our model modulates its initial estimate,  $\hat{A}_S$  to reflect different material configurations. For example, assigning *glass* or *carpet* to the floor, walls, and ceiling yields distinct patterns in our final RIR prediction,  $\hat{A}_M$ , indicating that variations in the material assignments in the environment influence our modeled scene acoustics. These qualitative results in the real-world setup demonstrate our model’s ability to produce material-aware RIRs in real-world settings.

### 7.8. Perceptual Material User Study

To evaluate the perceptual quality of our model predictions, we conduct a user study, where participants are required to rank two outputs—one from our approach and the other from the SOTA M-CAPA [62] baseline. Figure 10 shows the interface used for this study. For this user study, we select 22 random samples from our hardest test split,  $D_{uu}$ , where both the environments and the material configurations are not previously seen. Importantly, all qualitative examples shown in this paper are from this split. Having obtained the RIR predictions of a model for the  $D_{uu}$  split, we first convolve the RIRs with clean anechoic speech from the LibriSpeech [54]

dataset. Next, we provide an RGB image of a scene to the participants of the user study and ask them to **1)** first observe a certain target material mask and listen to the convolved audio produced using the ground-truth RIR corresponding to the spatial structure shown in the image and the given material mask, in order to get a sense of how the material properties affects the corresponding spatial audio; **2)** then listen to the spatial audios corresponding to our method and the M-CAPA baseline under a new material configuration but the same spatial layout; and **3)** finally rank the models vis-a-vis how well their predictions match the target materials. We observe that the participants preferred predictions from MatRIR over those from M-CAPA in 60.4% of the cases. This shows that our method can better capture changes in the material composition of the environment than previous work.

### 7.9. Evaluation setup

We evaluate our approach using existing methods and baselines that cover various aspects of RIR generation. *Image2Reverb* and *FAST-RIR++* are state-of-the-art (SoTA) approaches for RIR prediction using vision and spatial cues. Both these baselines exhibit low performance in evaluations, even after training with the AcoW dataset. This shows that reliance only on visual or spatial cues from the environment is not sufficient to generate accurate, material-conditioned RIRs.

Our approach disjointly models spatial and material cues from the environment, generating spatial-specific and material-specific RIRs for each material mapping and scene. To isolate the impact of this disentangled approach, we create *JM-\** baselines that jointly model all environmental cues using the same input features as our model. *JM-CNN* uses a CNN decoder approach to estimate the material-conditioned RIR. *JM-Transformer* uses a transformer encoder approach with the same architecture as our material encoder  $\mathcal{R}_M$ , to encode all environmental features jointly. *JM-QFormer* follows our spatial RIR decoder  $\mathcal{R}_S$  to learn a set of tokens by cross attending against input environmental features. We construct these baselines to explore the impact of joint modeling and compare out isolated, disentangled RIR prediction approach

Finally, we evaluate on current SoTA vision-based method that explicitly maps material configurations to RIRs, M-CAPA. This method uses visual input ( $V, M$ ) to directly estimate the material-conditioned RIR. Interestingly, our approach of separately predicting spatial and material RIRs produces more accurate results over SoTA methods that jointly model material-conditioned RIR.

## 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, 1979.
- [2] Lakulish Antani, Anish Chandak, Micah Taylor, and Dinesh Manocha. Direct-to-indirect acoustic radiance transfer. *IEEE Transactions on Visualization and Computer Graphics*, 18(2): 261–269, 2012.
  - [3] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, Jiankang Deng, and Xiatian Zhu. Av-gs: Learning material and geometry aware priors for novel view acoustic synthesis. In *Advances in Neural Information Processing Systems*, pages 28920–28937. Curran Associates, Inc., 2024.
  - [4] Reiner Birkel, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.
  - [5] Chunxiao Cao, Zhong Ren, Carl Schissler, Dinesh Manocha, and Kun Zhou. Interactive sound propagation with bidirectional path tracing. *ACM Trans. Graph.*, 35(6), 2016.
  - [6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.
  - [7] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicens Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020.
  - [8] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15516–15525, 2021.
  - [9] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868, 2022.
  - [10] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS 2022 Datasets and Benchmarks Track*, 2022.
  - [11] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6409–6419, 2023.
  - [12] Jiajian Chen, Jiakang Chen, Hang Chen, Qing Wang, Yu Gao, and Jun Du. Mean-rir: Multi-modal environment-aware network for robust room impulse response estimation, 2025.
  - [13] Mingfei Chen and Eli Shlizerman. Av-cloud: Spatial audio rendering through audio-visual cloud splatting. In *Advances in Neural Information Processing Systems*, pages 141021–141044. Curran Associates, Inc., 2024.
  - [14] Mingfei Chen, Israel D. Gebru, Ishwarya Ananthabhotla, Christian Richardt, Dejan Markovic, Jake Sandakly, Steven Krenn, Todd Keebler, Eli Shlizerman, and Alexander Richard. Soundvista: Novel-view ambient sound synthesis via visual-acoustic binding. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8331–8341, 2025.
  - [15] Chen, Changan and Ramos, Jordi and Tomar, Anshul and Grauman, Kristen. Sim2Real Transfer for Audio-Visual Navigation with Frequency-Adaptive Acoustic Field Prediction. In *IROS*, 2024.
  - [16] Sanjoy Chowdhury, Sreyan Ghosh, Subhrajyoti Dasgupta, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7884–7896, 2023.
  - [17] Orchisama Das, Paul Calamia, and Sebastia V. Amengual Gari. Room impulse response interpolation from a sparse set of measurements using a modal architecture. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964, 2021.
  - [18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.
  - [19] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B. Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707, 2020.
  - [20] Huiyu Gao, Jiahao Ma, David Ahmmedt-Aristizabal, Chuong Nguyen, and Miaomiao Liu. Soaf: Scene occlusion-aware neural acoustic field. *arXiv preprint arXiv:2407.02264*, 2024.
  - [21] Nail A. Gumerov and Ramani Duraiswami. A broadband fast multipole accelerated boundary element method for the three dimensional Helmholtz equation. *The Journal of the Acoustical Society of America*, 125(1):191–205, 2009.
  - [22] Brian Hamilton and Stefan Bilbao. Fdtd methods for 3-d room acoustics simulation with high-order accuracy in space and time. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2112–2124, 2017.
  - [23] Dorte Hammershøi and Henrik Møller. Binaural technique—basic methods for recording, synthesis, and reproduction. *Communication acoustics*, pages 223–254, 2005.
  - [24] Jun Han and Claudio Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks*, pages 195–201. Springer, 1995.
  - [25] F.J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.
  - [26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
  - [27] Martin Holters, Tobias Corbach, and Udo Zölzer. Impulse response measurement techniques and their applicability in the real world, 2009.
  - [28] Joanna Hong, Minsu Kim, Daehun Yoo, and Yong Man Ro. Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition. In *Interspeech*, 2022.

- [29] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [30] Derong Jin and Ruohan Gao. Differentiable room acoustic rendering with multi-view vision priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 37–47, 2025.
- [31] Derong Jin and Ruohan Gao. Differentiable room acoustic rendering with multi-view vision priors. In *International Conference on Computer Vision (ICCV)*, 2025.
- [32] Christian Kehling. Evaluation of data augmentation techniques of room impulse responses for improved ai-based estimations. In *Proc. DAGA*, pages 1285–1288, 2024.
- [33] Hansung Kim, Luca Remaggi, Philip JB Jackson, and Adrian Hilton. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 120–126. IEEE, 2019.
- [34] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [35] Homare Kon and Hideki Koike. Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images. In *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.
- [36] Homare Kon and Hideki Koike. An auditory scaling method for reverb synthesis from a single two-dimensional image. *Acoustical Science and Technology*, 41(4):675–685, 2020.
- [37] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.
- [38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023.
- [39] Tingle Li, Renhao Wang, Po-Yao Huang, Andrew Owens, and Gopala Anumanchipalli. Self-supervised audio-visual soundscape stylization. In *European Conference on Computer Vision*, pages 20–40. Springer, 2024.
- [40] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *ArXiv*, abs/2302.02088, 2023.
- [41] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields, 2023.
- [42] Susan Liang, Chao Huang, Yunlong Tang, Zeliang Zhang, and Chenliang Xu. p-avas: Can physics-integrated audio-visual modeling boost neural acoustic synthesis? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13942–13951, 2025.
- [43] Shiguang Liu and Dinesh Manocha. *Sound synthesis, propagation, and rendering*. Morgan & Claypool Publishers, 2022.
- [44] Xiulong Liu, Anurag Kumar, Paul Calamia, Sebastia V Amengual, Calvin Murdock, Ishwarya Ananthabhotla, Philip Robinson, Eli Shlizerman, Vamsi Krishna Ithapu, and Ruohan Gao. Hearing anywhere in any environment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5732–5741, 2025.
- [45] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [46] Andrew Luo, Yilun Du, Michael J. Tarr, Joshua B. Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. In *Advances in Neural Information Processing Systems*, 2022.
- [47] Sagnik Majumder and Kristen Grauman. Active audio-visual separation of dynamic sound sources. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 551–569. Springer, 2022.
- [48] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 275–285, 2021.
- [49] Sagnik Majumder, Changan Chen\*, Ziad Al-Halah\*, and Kristen Grauman. Few-Shot Audio-Visual Learning of Environment Acoustics. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- [50] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: theoretical analysis and applications. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023.
- [51] Ravish Mehra, Nikunj Raghuvanshi, Lauri Savioja, Ming C. Lin, and Dinesh Manocha. An efficient gpu-based time domain solver for the acoustic wave equation. *Applied Acoustics*, 73(2):83–94, 2012.
- [52] Shentong Mo and Yapeng Tian. Av-sam: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*, 2023.
- [53] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023.
- [54] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.
- [55] Nikunj Raghuvanshi, Rahul Narain, and Ming C. Lin. Efficient and accurate sound propagation using adaptive rectangular decomposition. *IEEE Transactions on Visualization and Computer Graphics*, 15(5):789–801, 2009.
- [56] Anton Ratnarajah and Dinesh Manocha. Listen2scene: Interactive material-aware binaural sound propagation for reconstructed 3d scenes. *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 254–264, 2023.
- [57] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. TS-RIR: Translated synthetic room impulse responses for speech augmentation. In *2021 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 259–266. IEEE, 2021.



- [58] 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.
- [59] Anton Ratnarajah, Ishwarya Ananthabhotla, Vamsi Krishna Ithapu, Pablo Hoffmann, Dinesh Manocha, and Paul Calamia. Towards improved room impulse response estimation for speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [60] 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, 2024.
- [61] Luca Remaggi, Hansung Kim, Philip JB Jackson, and Adrian Hilton. Reproducing real world acoustics in virtual reality using spherical cameras. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.
- [62] Mahnoor Fatima Saad and Ziad Al-Halah. How Would It Sound? Material-Controlled Multimodal Acoustic Profile Generation for Indoor Scenes. In *International Conference on Computer Vision (ICCV)*, 2025.
- [63] Carl Schissler and Dinesh Manocha. Interactive sound propagation and rendering for large multi-source scenes. *ACM Transactions on Graphics (TOG)*, 36(4):1, 2016.
- [64] Carl Schissler, Christian Loftin, and Dinesh Manocha. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE transactions on visualization and computer graphics*, 24(3):1246–1259, 2017.
- [65] Carl Schissler, Christian Loftin, and Dinesh Manocha. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE Transactions on Visualization and Computer Graphics*, 24(3):1246–1259, 2018.
- [66] Kristina P. Sinaga and Miin-Shen Yang. Unsupervised k-means clustering algorithm. *IEEE Access*, 8:80716–80727, 2020.
- [67] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 286–295, 2021.
- [68] Arjun Somayazulu, Changan Chen, and Kristen Grauman. Self-supervised visual acoustic matching. *Advances in Neural Information Processing Systems*, 36, 2024.
- [69] Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. Comparison of different impulse response measurement techniques. *Journal of the Audio Engineering Society*, 50(4):249–262, 2002.
- [70] Kun Su, Mingfei Chen, and Eli Shlizerman. INRAS: Implicit neural representation for audio scenes. In *Advances in Neural Information Processing Systems*, 2022.
- [71] Zhenyu Tang, Rohith Aralikatti, Anton Jeran Ratnarajah, and Dinesh Manocha. Gwa: A large high-quality acoustic dataset for audio processing. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.
- [72] Lonny L. Thompson. A review of finite-element methods for time-harmonic acoustics. *The Journal of the Acoustical Society of America*, 119(3):1315–1330, 2006.
- [73] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [74] Tor Erik Vigran. *Building acoustics*. CRC Press, 2014.
- [75] Michael 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, 1989.
- [76] Mason Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *CVPR*, 2024.
- [77] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022.
- [78] Stephan Werner, Florian Klein, Annika Neidhardt, Ulrike Sloma, Christian Schneiderwind, and Karlheinz Brandenburg. Creation of auditory augmented reality using a position-dynamic binaural synthesis system—technical components, psychoacoustic needs, and perceptual evaluation. *Applied Sciences*, 11(3):1150, 2021.
- [79] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53, 2013.
- [80] Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. Binaural audio-visual localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2961–2968, 2021.
- [81] Bing Xu. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.
- [82] Wen Zhang, Parasanga N Samarasinghe, Hanchi Chen, and Thushara D Abhayapala. Surround by sound: A review of spatial audio recording and reproduction. *Applied Sciences*, 7(5):532, 2017.