How Would It Sound?
Material-Controlled Multimodal Acoustic Profile Generation for Indoor Scenes

Supplementary Material

6. Supplementary Material

In this supplementary material, we provide further details about:

- Supplementary video (with audio) Sec. 6.1 for qualitative evaluation of our model predictions as stated in Sec. 4.
- Real-World Generalization (Sec. 6.3) as mentioned in Sec. 4.
- Ablations on other test splits (Sec. 6.2) as mentioned in Sec. 4, Table 2.
- Loss ablations Sec. 6.4 and computational cost analysis Sec. 6.5 as stated in Sec. 4.
- Performance analysis on other test splits (Sec. 6.6) as stated in Sec. 4.
- Evaluation results on seen splits (Sec. 6.7) as stated in (Sec. 4).
- Robustness to noise experiments (Sec. 6.8) as noted in Sec. 4.
- Acoustic Wonderland dataset (Sec. 6.9), as mentioned in Sec. 3.2, and a user study on the perceptual differences as mentioned in Sec. 3.2.
- Model Architecture details (Sec. 6.10).
- Evaluation setup (Sec. 6.11), as mentioned in Sec. 4.

6.1. Supplementary Video

We provide a supplementary video, see the project page, to illustrate the qualitative results produced by our model, M-CAPA. The video begins with a brief overview of the motivation and contributions of this work. It then presents qualitative results by showcasing a variety of speech sounds from the datasets [39] and [30], convolved with the predicted target room impulse response (RIR), $\hat{A}_T$. These examples emphasize the quality of the predictions and demonstrate how effectively the model captures the diverse target material configurations introduced in the input scenes.

Furthermore, the video highlights failure cases where the model encountered difficulties in accurately representing material changes, thereby shedding light on challenges that remain to be addressed. For instance, M-CAPA struggles to model environmental acoustics when significant material changes are applied to large objects with highly irregular shapes. Additionally, we observe suboptimal performance when certain materials, such as *Sound-Proof* and *Steel*, are extensively used in the target material mask.

6.2. Ablations On Other Test Splits

We present ablation results on the remaining test splits, D_{us} and D_{uk}, in Table 3. Similar trends to those reported in Table 2 in the main text are observed. Our complete model, *M-CAPA*, achieves the best overall performance across all splits. Notably, as shown in *row b*, incorporating B_T allows the model to learn the differences between A_S and A_T that arise from selecting target materials, which introduce new types of reverberations not present in A_S. This incorporation enhances learning, particularly for acoustic metrics such as RTE and CTE. Furthermore, in *row a*, excluding the target material change and relying solely on visual cues and A_S to predict A_T leads to a noticeable degradation in performance.

	Unseen Environments							
	Seen Materials				Unseen Pairings			
Method	L1	STFT	RTE	CTE	L1	STFT	RTE	CTE
M-CAPA (Ours)	5.10	3.62	88.15	8.04	5.47	4.15	91.32	8.57
a) Ours w/o $\mathcal{M}_T$	5.39	3.78	104.77	8.67	5.77	4.35	107.53	9.13
b) Ours w/o B_T	5.52	4.52	98.30	10.79	5.93	5.17	104.72	10.51
c) Ours w/ Inferred G_n	5.42	3.72	98.46	8.53	5.79	4.27	99.70	9.03
d) Ours w/ Changed $\mathcal{M}_T$	5.27	3.74	94.97	8.48	5.63	4.29	96.81	8.95

Table 3. Ablation results of our model on unseen environments using test sets D_{us} (seen material profiles) and D_{uk} (unseen material profile pairings). The results exhibit similar trends to those observed on D_{uu}. For all metrics, lower values indicate better performance.

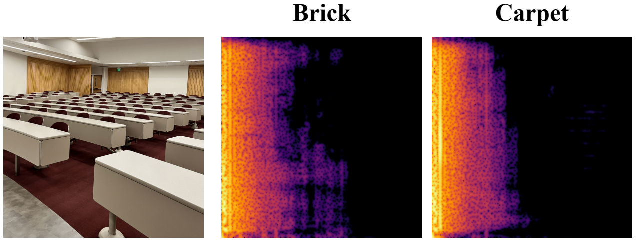


Figure 6. Predicted RIRs from vision-only M-CAPA in an auditorium classroom environment where M_T=Brick and M_T=Carpet

6.3. Real-World Generalization

To asses M-CAPA's ability to generalize to real-world samples, we collected RGB images from two real-world scenes and used our vision-only M-CAPA to generate a target RIR (A_T). The target material of the objects in the scenes was set to one of three classes *carpet, brick, and glass* (Figure 6 shows qualitative results).

Then, we conduct a user study (4.3) to measure M-CAPA's performance. We ask 5 users to go through a brief training so they may distinguish the acoustic properties of different materials (Figure 7a). Afterwards, we ask them

Loss	L1	STFT	RTE	CTE
$L_1 + L_2$+Energy Decay	5.29	3.87	90.61	8.52
a) L_1 Only	5.46	4.13	97.92	9.47
b) L_2 Only	6.19	4.00	241.41	9.22
c) L_1+Energy Decay	5.55	4.15	99.00	9.45
d) L_2+Energy Decay	6.47	4.12	248.69	9.12
e) $L_1 + L_2$	5.59	3.99	109.27	9.26

Table 4. Ablation of losses

Method	A_S	V_n	Params (M)	GFLOPs	Inf. Time (ms)
AV-RIR [37]	✓	✓	390.66	270.43	794.06
M-CAPA (Ours)	✓	✓	**10.56**	**17.98**	**114.22**
Image2Reverb [44]		✓	57.6	276.91	198.44
FAST-RIR[35]++		✓	132.68	57.84	121.76
M-CAPA (Ours)		✓	**5.84**	**11.24**	**76.61**

Table 5. Computational cost of the baselines and M-CAPA. Our approach is significantly faster and lighter. Lower is better for all metrics.

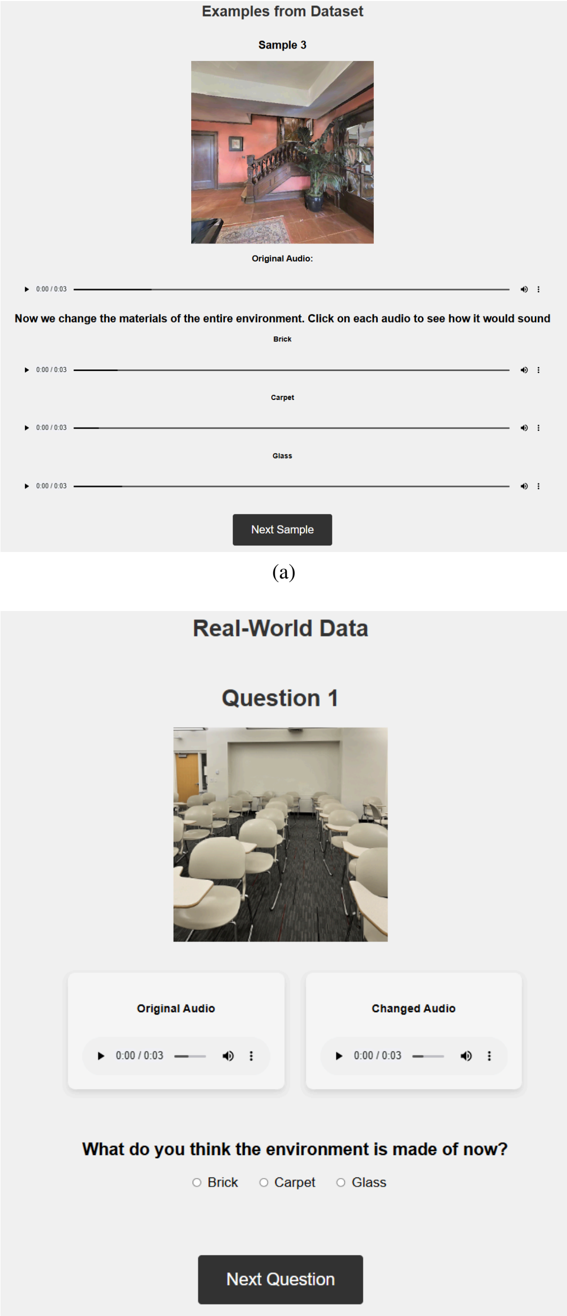
to listen to the predictions by M-CAPA on the real-world samples when A_T is convolved with speech, and ask them to identify the target material used to generate A_T as one of the three materials: Brick, Carpet, and Glass (Figure 7b). Overall, the accuracy achieved by the users in identifying the correct material in this task was 61.1% (random chance: 33%), showing that our model successfully encodes the target material signature in A_T even in samples from real-world scenes.

6.4. Loss Ablations

As discussed in Sec.3.3, our model is trained with $L1$, $L2$ and energy decay loss [27]. We investigate the impact of each loss as our learning objective by performing ablations on the losses (Table 4). We see from row (a) and row (b) that $L1$ is the most important loss in minimizing error between predicted RIR and ground truth RIR. However, $L2$ plays a vital role in ensuring that the STFT loss is minimized, and that loss between acoustic parameters is consequently reduced. The energy decay loss acts as supervision for the acoustic metrics, CTE and RTE, ensuring that the reverberation time and early-to-late reflections of the predicted RIR are aligned with the ground truth RIR.

6.5. Computational Cost

Our M-CAPA is a light-weight and efficient end-to-end model that can render RIRs conditioned on material profiles. Table 5 compares the number of trainable parameters, GFLOPs, and inference time of M-CAPA to other SoTA approaches. Our model is significantly faster and lighter than the baselines.


Examples from Dataset

Sample 3

Now we change the materials of the entire environment. Click on each audio to see how it would sound

(a)

Real-World Data

Question 1

What do you think the environment is made of now?

○ Brick ○ Carpet ○ Glass

(b)

Figure 7. User interface for the real-world user study. a) Interface for user training b) Interface for the real-world samples.

6.6. Performance Analysis on D_{us} and D_{uk}

We analyze the performance of our model with respect to the changed material area in $\mathcal{M}_T$ and the different material classes, on the remaining test splits D_{us} (Fig. 8) and D_{uk} (Fig. 9). In both cases, we observe that our model generally benefits from material changes applied to larger areas within the scene. Larger areas provide more information to the model about how the target acoustic profile may change,

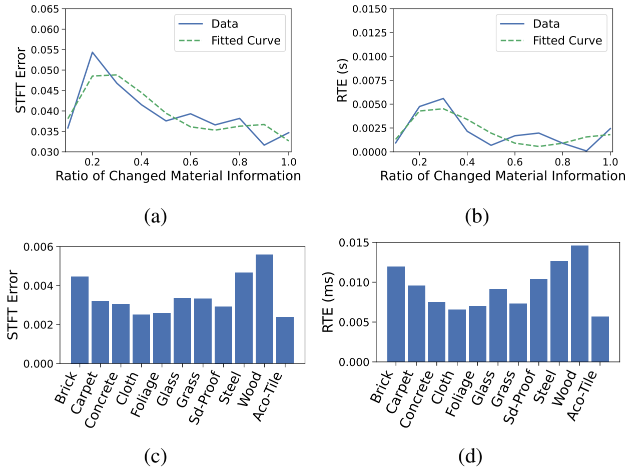


Figure 8. Performance analysis of our model on D_{us} with respect to the percentage of new material assignments in $\mathcal{M}_T$ (a and b) and across different material classes (c and d).

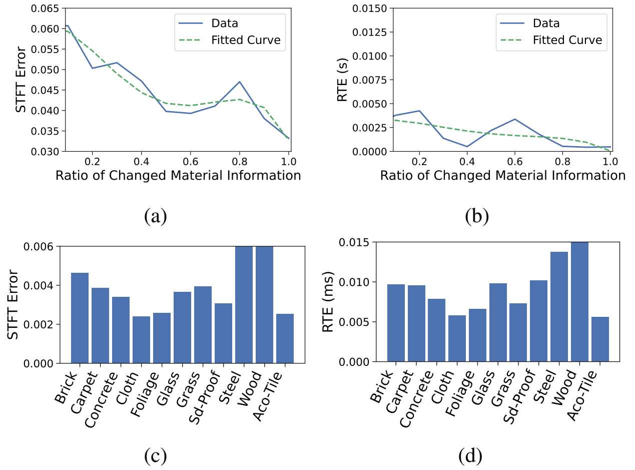


Figure 9. Performance analysis of our model on D_{uk} with respect to the percentage of new material assignments in $\mathcal{M}_T$ (a and b) and across different material classes (c and d).

compared to cases where only a small area undergoes new material assignments.

Furthermore, consistent with our analysis of performance on D_{uu}, we find that certain material classes, such as *Steel* and *Wood*, are relatively more challenging for the model to accurately predict compared to others.

6.7. Evaluation Results on Seen Environments

			Seen Environments							
	Observation		**Seen Materials** (D_{ss})				**Unseen Materials** (D_{su})			
Method	A_s	V_n	**L1**	**STFT**	**RTE**	**CTE**	**L1**	**STFT**	**RTE**	**CTE**
Direct Mapping	✓		8.22	8.29	121.01	12.07	8.33	8.27	120.97	12.99
M-CAPA (Ours)	✓		**5.96**	**4.63**	**92.33**	**7.73**	**5.98**	**4.62**	**93.96**	**8.72**
Image2Reverb[44]		✓	14.35	7.60	253.02	20.95	14.12	7.39	237.69	21.48
FAST-RIR++[27, 35]		✓	17.25	32.45	303.95	22.95	17.21	33.51	316.15	21.91
Material Agnostic		✓	8.18	8.11	119.23	11.47	8.23	8.24	117.03	12.33
Material Aware		✓	**3.47**	**3.36**	**57.68**	**5.09**	7.27	7.02	**83.91**	9.79
M-CAPA (Ours)		✓	5.98	5.17	90.16	7.62	**5.96**	**5.05**	91.59	**8.64**
AV-RIR [37]	✓	✓	7.66	8.14	**64.47**	10.56	8.16	8.22	**85.83**	11.67
M-CAPA (Ours)	✓	✓	**5.80**	**4.63**	90.72	**7.71**	**5.81**	**4.61**	91.56	**8.70**

Table 6. Results on seen environments (used during training) when evaluated under two conditions: when coupled with seen material profiles (D_{ss}) which match exactly the training setup, and when coupled with unseen material profiles (D_{su}). Certain methods, such as the Material Aware, appear to overfit to the training samples in D_{ss}, leading to poor generalization performance on unseen cases like those in D_{su}. In contrast, our model, M-CAPA, demonstrates better generalization capabilities, achieving improved performance on D_{su} while maintaining balanced results on D_{ss}. STFT and L_1 are scaled by $\times 10^{-2}$, RTE is in milliseconds (ms), and CTE in decibels (dB). Lower values indicate better performance for all metrics.

We present the performance of our model in *seen* environments in Table 6. These environments are observed during training, and we evaluate performance under two setups: with seen material profiles (D_{ss}) and with unseen material profiles (D_{su}). The results for the split where both environments and materials match the training setup (D_{ss}) show that baselines, such as the Material Aware baseline, perform exceptionally well. This is expected, as both the evaluation and training samples originate from the same scene and material distributions, enabling these baselines to overfit effectively to the training data. However, this overfitting results in poor generalization to unseen material profiles (D_{su}), as shown in the left side of Table 6, and limited generalization to unseen environments, as highlighted in the main experiments (Table 1). In contrast, our model, *M-CAPA*, demonstrates robust generalization across unseen material profiles and unseen environments, as demonstrated by the results.

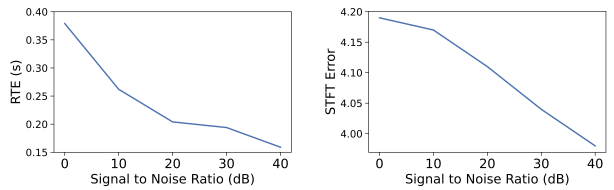
6.8. Noise Experiments

We evaluate the robustness of our model against noisy estimates of A_S. During inference, we introduce Gaussian noise to the source RIR with varying levels of strength, ranging from a signal-to-noise ratio (SNR) of 40 dB (relatively clean A_S) to 0 dB (extremely noisy A_S). In Fig. 10, we illustrate the impact of noise on our model's performance for both the STFT error and the RTE metrics on the D_{uu} split (a similar trend is observed on the other test splits).

Our results show that the model's performance degrades gradually as the noise level increases. We believe that the robustness of our model to noise could be improved by incorporating data augmentation techniques with noisy inputs during training. We leave this as a direction for future work.

6.9. Acoustic Wonderland Dataset

We provide detailed information regarding the creation and characteristics of our dataset, including the location sampling methodology, material properties, material profiles, and their pairings.



(a) (b)

Figure 10. Robustness to noise. We introduce increasing levels of noise to the source RIR A_S during inference, ranging from an SNR of 40 dB (clean A_S) to 0 dB (extremely noisy A_S), and evaluate performance on D_{uu}. For both metrics, lower values indicate better performance.

Location Sampling The locations for sampling data points in our dataset are selected based on specific criteria to ensure that each point lies in an open space within the environment and provides meaningful visual and acoustic information. The sampling process involves randomly selecting locations within an indoor scene, subject to the condition that no two sampled locations are closer than a predefined distance threshold of *0.1m*. This prevents sampling overlapping locations and ensures a more uniform spatial coverage of the scene. At each selected location, we place a sensor suite consisting of a camera, a speaker, and binaural microphones with a random orientation. To enhance the diversity and realism of the dataset, care is taken to avoid situations where the camera is positioned too close to, or directly facing, large objects such as walls or doors.

Material Classes Our dataset incorporates 12 material classes, including *wood*, *steel*, *concrete*, *grass*, *foliage*, *glass*, *brick*, *steel*, *sound-proof*, *carpet* and *acoustic tiles*. We also include a *default* material class which is SoundSpaces default material mapped onto any unlabeled object in the scene. Each material class is characterized in SoundSpaces by its acoustic coefficients, such as reflection, absorption, transmission, and damping properties across various frequency bands of sound waves. These coefficients are essential for accurately modeling the acoustic behavior of the materials within the simulated environment.

Material Profiles Each profile defines a mapping between material classes and semantic object categories within a scene. The SoundSpaces simulator utilizes this mapping to assign materials to objects based on their semantic labels. For each material profile in our dataset, a random mapping is generated to disentangle the relationship between material and semantic classes. For instance, one material profile may assign *wall* and *floor* to the material *wood*, while another profile maps *wall* to *concrete* and *floor* to *carpet*. These mappings are applied to large objects and surfaces, such as furniture, doors, and walls, while smaller objects (e.g., sports equipment, utensils, televisions) retain their default materials. This distinction is made because smaller objects typically have negligible impact on the overall acoustic profile of the scene. In total, we generate *2,673* unique material profiles for our dataset. See examples in Fig. 11.

Pairings Following the observation sampling step described in the main paper (Sec. 3.2), we sample, for each location, a random pairing of two observations derived from different material profiles: $O_{n,S} = (V_n, G_n, \mathcal{M}_{n,S}, A_{n,S})$ and $O_{n,T} = (V_n, G_n, \mathcal{M}_{n,T}, A_{n,T})$. In this pairing, one observation serves as the source configuration, representing the original state of the scene $(V_n, G_n, A_{n,S})$, while the other represents the target state $(\mathcal{M}_{n,T}, A_{n,T})$, after applying a material change. The material change is denoted as $diff(\mathcal{M}_{n,T}, \mathcal{M}_{n,S})$. This setup simulates a scenario where a user alters the material configuration of the scene from $\mathcal{M}_{n,S}$ to $\mathcal{M}_{n,T}$, and the objective is to generate the corresponding target acoustic profile $A_{n,T}$.

Perceptual Differences When collecting our dataset, we filtered out any samples in which less than 10% of the input view contained changed material to ensure a noticeable difference between A_S and A_T. However, does our data correspond to samples with noticeable perceptual differences observed by the users? To investigate this, we selected 45 samples uniformly from various L_2 differences between A_S and A_T in our test data, along with 15 controlled samples featuring identical RIR pairs where $A_S = A_T$. We then asked 8 users to listen to sounds convolved with both RIRs and determine whether they sounded the same or different.

Our results show that the users achieved 87.9% accuracy, indicating a strong perceptual distinction in our dataset. We show error distribution for the user study in Figure 12b. Most errors occurred when the L_2 difference was in the lower range (11.1 to 77.8), suggesting that smaller variations in L2 distance are less perceptually salient. However, in general the error is low, below 6%, across all L_2 bins.

In Table 7 and Table8, we present the performance of different models on our test data, focusing only on samples with high perceptual differences ($L_2 \geq 75$). The results show that our model maintains its advantage over state-of-the-art and baselines in this setting as well.

6.10. Model Architecture Details

The encoders in our model are based on a convolutional neural architecture inspired by the UNet [40]. Each encoder (f^V, f^G, f^A, or f^M) comprises four downsampling layers. Each layer includes a convolutional block followed by a downsampling module.

The convolutional block consists of two consecutive Conv2D layers, each with a kernel size of 3, a batch normalization layer, and a LeakyReLU activation [57]. To enhance generalization, a dropout layer [46] with a rate of 0.2 is included in each layer.

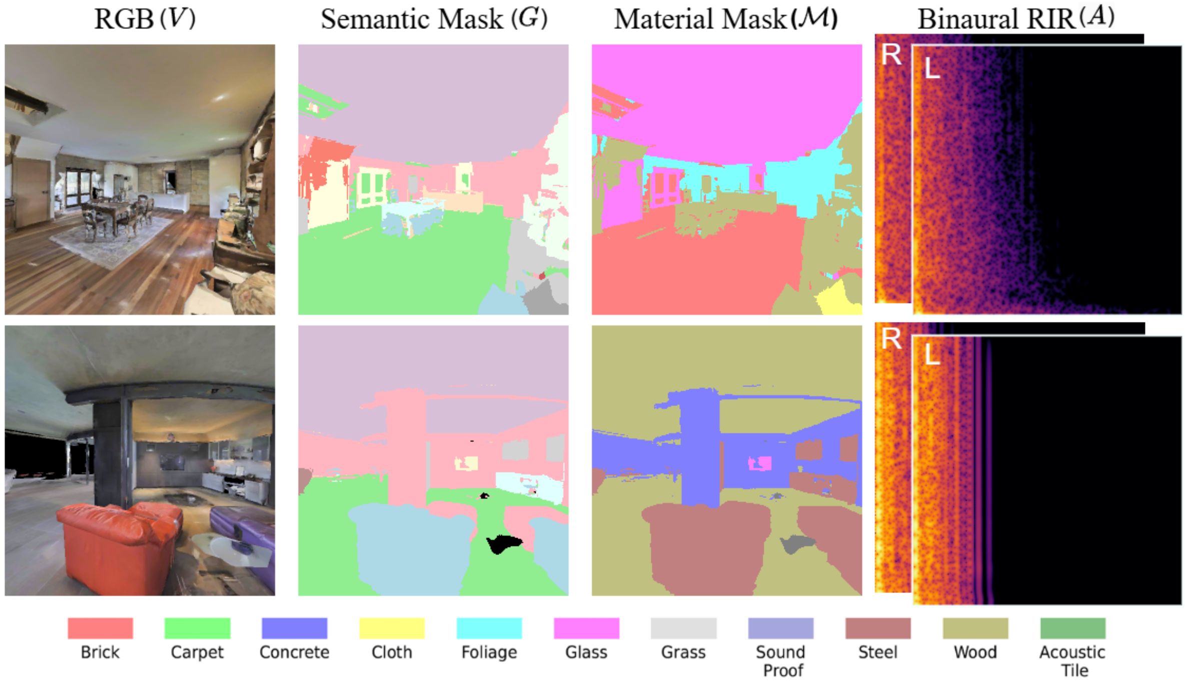


Figure 11. Examples from our Acoustic Wonderland Dataset. Each data point contains an RGB image, a semantic segmentation mask, a material segmentation mask, and the corresponding acoustic profile in the form of a two-channel RIR.

	Observation		**Seen Materials**				**Unseen Materials**				**Unseen Pairings**			
Method	A_S	V_n	**L1**	**STFT**	**RTE**	**CTE**	**L1**	**STFT**	**RTE**	**CTE**	**L1**	**STFT**	**RTE**	**CTE**
Direct Mapping	✓		9.63	10.29	132.7	14.65	9.59	10.31	134.4	15.04	9.97	10.89	133.9	14.03
M-CAPA (Ours)	✓		**6.75**	**5.38**	**98.28**	**9.05**	**6.76**	**5.42**	**102.2**	**9.41**	**7.25**	**6.12**	**100.2**	**9.91**
Image2Reverb [44]		✓	18.38	9.51	234.1	39.92	17.56	8.91	202.2	40.65	16.36	9.27	231.5	37.89
FAST-RIR++ [27, 35]		✓	18.97	34.88	311.4	20.78	18.67	37.29	324.8	20.30	19.71	44.80	312.0	20.67
Material Agnostic		✓	10.06	13.27	127.8	14.28	10.12	13.01	127.1	14.56	10.49	13.76	129.9	13.93
Material Aware		✓	9.88	12.64	105.2	11.81	9.81	12.65	102.5	12.18	10.60	13.75	106.3	12.05
M-CAPA (Ours)		✓	**7.16**	**7.23**	**96.90**	**9.30**	**7.13**	**7.23**	**98.28**	**9.65**	**7.70**	**8.24**	**101.3**	**10.03**
AV-RIR [37]	✓	✓	9.62	10.30	108.3	12.78	9.57	10.32	106.7	12.78	10.06	10.97	107.4	12.35
M-CAPA (Ours)	✓	✓	**6.57**	**5.39**	**97.58**	**8.99**	**6.54**	**5.42**	**101.0**	**9.22**	**7.07**	**6.15**	**101.6**	**9.81**

Table 7. Results on unseen environments with (A_S, A_T) samples that have $L_2 \geq 75$ for our three test splits: D_{us}, D_{uu} and D_{uk}. STFT and L_1 are scaled by $\times 10^{-2}$, RTE is in milliseconds (ms), and CTE in decibels (dB). Lower values indicate better performance for all metrics.

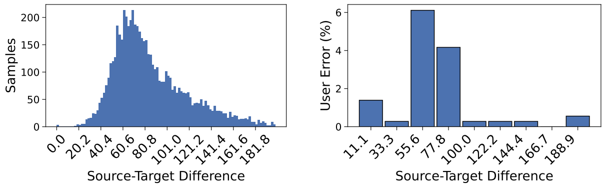
The downsampling module within each encoder layer consists of a MaxPooling layer with a kernel size of 2 and a stride of 2. This reduces the spatial resolution by a factor of 2 at each layer. The four layers of the encoder use 32, 64, 128, and 512 kernels, respectively.

The fusion layer, $\mathcal{F}$, combines the multimodal scene embedding e_m and the material embedding e_t. This fusion is performed using a single Conv2D layer with a kernel size of 3 and a stride of 1, which effectively integrates information from both embeddings into a unified representation.

The decoder, f^T, follows an architecture similar to the encoders but in a mirrored configuration. It consists of four upsampling blocks. Each upsampling block contains a single Transpose Conv2D layer, followed by two Conv2D layers, a batch normalization layer, and a LeakyReLU activation function. Skip connections are incorporated from the corresponding layers of the f^A encoder, allowing the decoder to leverage features from earlier stages of the encoding process. The final output of the decoder is a two-channel binaural magnitude spectrogram of the target acoustic response.

Method	L1	STFT	RTE	CTE
M-CAPA (Ours)	6.56	5.42	101.0	9.25
a) Ours w/o $\mathcal{M}_T$	6.91	5.64	117.1	10.02
b) Ours w/o B_T	7.20	6.98	116.3	12.52
c) Ours w/ Inferred G_n	6.93	5.56	107.3	9.96
d) Ours w/ Changed $\mathcal{M}_T$	6.78	5.59	108.1	9.91

Table 8. Ablation of our model on the test split D_{uu} with distance between $(A_S, A_T) \geq 75$. Lower is better for all metrics.



(a)

(b)

Figure 12. Analysis of perceptual differences in test data. Left, we show the distribution of differences between (A_S, A_T) in our unseen environments test splits. Right, we analyze the breakdown of errors accumulated by users during the perceptual difference user study. Overall, the error is low across all bins (below 6%), and as the L_2 distance between A_S and A_T increases, perceptual differences become more apparent and user error decreases.

6.11. Evaluation Setup

In this section, we provide additional details about the baselines and evaluation metrics used in our experiments.

Baselines

- **Direct Mapping**: This baseline directly uses A_S as the prediction for A_T, effectively ignoring the target material information. In other words, it assumes that the original acoustic response is sufficient to predict the target response. This baseline serves as a reference for quantifying the impact of material configuration on the target acoustics, as A_S already captures the scene shape, object distribution, and original material configuration.
- **Material Agnostic Matcher**: In this baseline, we compute the cosine similarity between the visual embedding e_v of the input and the embeddings of visual observations V_n in the training set. The most similar data point is selected, and a random RIR associated with that location l_n is returned as the prediction. This approach represents methods that estimate RIRs based on visual characteristics of the scene alone, without incorporating material information.
- **Material Aware Matcher**: Similar to the Material Agnostic Matcher, this baseline identifies the most visually similar scene location l_n from the training data. However, in addition to visual similarity, it takes material information into account. From the set of RIRs associated with different material profiles at the selected location, we compute the L1 distance between the material distribution associated with each RIR and the target material distribution $\mathcal{M}_T$. The RIR with the most similar material distribution to $\mathcal{M}_T$ is selected. This baseline highlights the importance of accounting for material configuration and the similarity between material settings during training and testing.
- **Image2Reverb** [44]: We follow the official implementation provided by the authors to train this model on our dataset. With the same pre-trained depth and visual encoders from the original implementation, we train the GAN-based network to predict RIRs using the Acoustic Wonderland dataset.
- **AV-RIR** [37]: The AV-RIR model initially infers the RIR from reverberant speech and then estimates the late components of the RIR using a retrieved sample from a material-aware training database. To adapt this baseline to our case and improve its performance, we make the following changes: (1) Instead of inferring the source RIR from reverberant speech, we provide A_S directly as input, as it offers a more accurate representation; (2) Similar to the Material Aware Matcher baseline, we retrieve the RIR of the closest training sample based on both visual and material-based similarity to the input sample. (3) While the original implementation uses a $\mathbf{360^\circ}$ panoramic RGB images to predict target RIRs, we choose to retrieve the closest sample in the training set using $\mathbf{90^\circ}$ Field of View (FoV) for fair comparison with M-CAPA which also uses $\mathbf{90^\circ}$ FoV. When comparing the impact of FoV on the performance of the AV-RIR baseline, we note that an increased FoV yields only marginal improvement. For example, in test split D_{uu}, L1 error drops from 7.59 to 7.49, STFT error reduces from 7.17 to 7.12, RTE improves from 99.10ms to 98.56ms and CTE drops from 11.35 to 11.22. This suggests that A_S already carries significant cues about the entire room, without needing $\mathbf{360^\circ}$ FoV as visual input. Following the AV-RIR approach, we retain the first 2000 samples of A_S and replace the remaining samples with the reverberant components of the retrieved RIR.
- **FAST-RIR++**: [35] is a GAN-based approach to RIR synthesis for rectangular rooms, using properties of the acoustic environment such as room size, speaker/listener positions and reverberation time of the target RIR. We modify this approach following [27] by making the following changes: (1) Instead of providing the room size, we provide ground truth depth images, making this a vision-based variation of the original implementation. (2) In addition to RT60 provided by the original implementation, we also provide the direction-to-reverberant ratio (DRR) as an acoustic parameter of the room. We obtain acoustic parameters from the source RIR. We train FAST-RIR++ on our training dataset until convergence and evaluate on test splits.

These baselines and existing methods address various aspects of evaluation and represent key directions in the RIR prediction literature. The *Direct Mapping* baseline evaluates methods that focus solely on capturing the geometric and structural properties of the scene, without accounting for material changes. In contrast, the *Material Agnostic* and *Material Aware* baselines represent robust nearest-neighbor approaches. These baselines rely on the similarity between test and training scenes, either based purely on visual information or incorporating material representations. This comparison enables us to evaluate whether a method merely memorizes training data and whether the inclusion of material information leads to improved predictive performance.

Furthermore, *Image2Reverb*, *FAST-RIR*++, and *AV-RIR* represent state-of-the-art (SoTA) approaches for RIR prediction. *Image2Reverb* relies exclusively on visual inputs to predict the RIR of a scene. Interestingly, our findings reveal that *Image2Reverb* demonstrates low performance in evaluations, even after retraining on our dataset, being outperformed by some of the baselines in RTE and CTE. This observation shows that reliance on just RGB observations is not sufficient to render accurate RIRs that model material changes in the environment. *AV-RIR* integrates material information within a more advanced prediction framework, estimating RIRs from reverberant speech, and finally conditioning late components of the estimated RIR using scene-based retrieval. AV-RIR focuses on limited material-object mapping, while our approach assumes all semantic objects in the scene are mapped to materials and contribute to the final RIR prediction. *FAST-RIR*++ provides an acoustically guided approach to RIR prediction, using target acoustic parameters to guide RIR generation. This baseline examines the impact of explicit acoustic parameters for the prediction of accurate RIRs.

Metrics We used the following metrics to evaluate performance:

- **L1 Error**: The L1 norm between the generated $\hat{A}_T$ and ground truth A_T audio's magnitude spectrograms.
- **STFT Error**: The mean squared error (MSE) between the magnitude spectrograms of the generated and ground truth audio's magnitude spectrograms.
- **RTE**: This metric (Reverberation Time Error) quantifies the difference in time taken for the energy of the predicted signal $\hat{A}_T$ and the ground truth signal A_T to decay by 60 dB. This is a standard metric used in prior works (e.g., [11, 27, 37]). Following the approach in [27], we use the Schroeder Integration Method [21] to estimate the decay time. For binaural RIRs, we compute the reverberation time for both channels and report the average absolute difference between $\hat{A}_T$ and A_T.
- **CTE [53]**: This metric calculates the difference in the ratio of direct energy (the first 50 ms of the signal) to late energy for both signals, providing insight into how accurately a model captures the acoustic characteristics of the environment.

Signal Reconstruction For both RTE and CTE, a waveform representation of $\hat{A}_T$ is required. Reconstructing the target signal accurately necessitates the inclusion of phase information. To address this, we leverage the phase information from the source impulse response (A_S). By carrying over the phase from A_S, the predicted magnitude can be reconstructed into a waveform that can be directly compared to the target waveform, ensuring a meaningful evaluation of the reconstruction accuracy.

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

[2] Lakulish Antani, Anish Chandak, Micah Taylor, and Dinesh Manocha. Direct-to-Indirect Acoustic Radiance Transfer. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 18(2):261–269, 2012. 2

[3] Chunxiao Cao, Zhong Ren, Carl Schissler, Dinesh Manocha, and Kun Zhou. Interactive sound propagation with bidirectional path tracing. *ACM Transactions on Graphics (TOG)*, 35(6), 2016. 2

[4] 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. 3

[5] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-Visual Navigaton in 3D Environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[6] 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. 2

[7] 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 (CVPR)*, pages 18858–18868, 2022. 3

[8] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning. In *NeurIPS Datasets and Benchmarks Track*, 2022. 2, 3

[9] 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 (CVPR)*, pages 6409–6419, 2023. 3

[10] Chen, Changan and Ramos, Jordi and Tomar, Anshul and Grauman, Kristen. Sim2Real Transfer for Audio-Visual Navigation with Frequency-Adaptive Acoustic Field Prediction. In

The IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2024. 3

[11] 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 (ICCV)*, pages 7884–7896, 2023. 2, 5, 7

[12] Orchisama Das, Paul Calamia, and Sebastia V. Amengual Gari. Room Impulse Response Interpolation from a Sparse Set of Measurements Using a Modal Architecture. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964, 2021. 2

[13] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B. Tenenbaum. Look, Listen, and Act: Towards Audio-Visual Embodied Navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707, 2020. 2

[14] 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. 2

[15] 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. 2

[16] 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. 5

[17] 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 *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 120–126. IEEE, 2019. 2

[18] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. 5

[19] 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. 3

[20] 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. 3

[21] Tobias Lentz, Dirk Schröder, Michael Vorländer, and Ingo Assenmacher. Virtual reality system with integrated sound field simulation and reproduction. *EURASIP journal on advances in signal processing*, 2007:1–19, 2007. 7

[22] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 3

[23] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. AV-NeRF: Learning Neural Fields for Real-World Audio-Visual Scene Synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2

[24] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural Acoustic Context Field: Rendering Realistic Room Impulse Response With Neural Fields, 2023. 2, 3

[25] Shiguang Liu and Dinesh Manocha. *Sound synthesis, propagation, and rendering*. Morgan & Claypool Publishers, 2022. 2

[26] 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 (NeurIPS)*, 2022. 2, 3

[27] 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. 3, 5, 6, 2, 7

[28] 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. 2

[29] Shentong Mo and Yapeng Tian. AV-SAM: Segment Anything Model Meets Audio-Visual Localization and Segmentation. *arXiv preprint arXiv:2305.01836*, 2023. 2

[30] 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. 1

[31] Nikunj Raghuvanshi, Rahul Narain, and Ming C. Lin. Efficient and Accurate Sound Propagation Using Adaptive Rectangular Decomposition. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 15(5):789–801, 2009. 2

[32] Anton Ratnarajah and Dinesh Manocha. Listen2Scene: Interactive material-aware binaural sound propagation for reconstructed 3D scenes . In *IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 254–264, 2024. 2, 3

[33] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. IR-GAN: Room Impulse Response Generator for Far-Field Speech Recognition. In *Proceedings of Interspeech 2021*, pages 286–290, 2021. 2

[34] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. TS-RIR: Translated Synthetic Room Impulse Responses for Speech Augmentation. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 259–266, 2021. 2, 6

[35] Anton Ratnarajah, Shi-Xiong Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. Fast-RIR: Fast Neural Diffuse Room Impulse Response Generator. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 571–575, 2022. 2, 3, 6, 5

[36] Anton Ratnarajah, Ishwarya Ananthabhotla, Vamsi Krishna Ithapu, Pablo Hoffmann, Dinesh Manocha, and Paul Calamia. Towards improved room impulse response estimation for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2

[37] 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. 2, 3, 6, 5, 7

[38] Luca Remaggi, Hansung Kim, Philip JB Jackson, and Adrian Hilton. Reproducing real world acoustics in virtual reality

using spherical cameras. In *International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019. 3

[39] Colleen Richey, Maria A Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, et al. Voices Obscured in Complex Environmental Settings (VOICES) corpus. *arXiv preprint arXiv:1804.05053*, 2018. 1

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. 4

[41] Carl Schissler and Dinesh Manocha. Interactive Sound Propagation and Rendering for Large Multi-Source Scenes. *ACM Transactions on Graphics (TOG)*, 36(4):1, 2016. 2, 3

[42] 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 (TVCG)*, 24(3):1246–1259, 2017. 3

[43] 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 (TVCG)*, 24(3):1246–1259, 2018. 2

[44] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 286–295, 2021. 2, 3, 6, 5

[45] Arjun Somayazulu, Changan Chen, and Kristen Grauman. Self-Supervised Visual Acoustic Matching. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 3

[46] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 4

[47] Kun Su, Mingfei Chen, and Eli Shlizerman. INRAS: Implicit Neural Representation for Audio Scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3

[48] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3

[49] Zhenyu Tang, Nicholas J. Bryan, Dingzeyu Li, Timothy R. Langlois, and Dinesh Manocha. Scene-aware audio rendering via deep acoustic analysis. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 26:1991–2001, 2019. 2

[50] Zhenyu Tang, Rohith Aralikatti, Anton Jeran Ratnarajah, and Dinesh Manocha. GWA: A large high-quality acoustic dataset for audio processing. In *Proceedings of the ACM Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 1–9, 2022. 3

[51] 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. 2

[52] 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. 2

[53] Tor Erik Vigran. *Building acoustics*. CRC Press, 2014. 1, 7

[54] 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. 2

[55] 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. 1

[56] 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. 2

[57] Bing Xu. Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv preprint arXiv:1505.00853*, 2015. 4

[58] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17830–17839, 2023. 5, 7