Learning Neural Acoustic Fields

Andrew Luo
Carnegie Mellon University

Yilun Du
Massachusetts Institute of Technology

Michael J. Tarr
Carnegie Mellon University

Joshua B. Tenenbaum
MIT BCS, CBMM, CSAIL

Antonio Torralba
Massachusetts Institute of Technology

Chuang Gan
UMass Amherst and MIT-IBM Watson AI Lab

Abstract

Our environment is filled with rich and dynamic acoustic information. When we walk into a cathedral, the reverberations as much as appearance inform us of the sanctuary's wide open space. Similarly, as an object moves around us, we expect the sound emitted to also exhibit this movement. While recent advances in learned implicit functions have led to increasingly higher quality representations of the visual world, there have not been commensurate advances in learning spatial auditory representations. To address this gap, we introduce Neural Acoustic Fields (NAFs), an implicit representation that captures how sounds propagate in a physical scene. By modeling acoustic propagation in a scene as a linear time-invariant system, NAFs learn to continuously map all emitter and listener location pairs to a neural impulse response function that can then be applied to arbitrary sounds. We demonstrate NAFs on both synthetic and real data, and show that the continuous nature of NAFs enables us to render spatial acoustics for a listener at arbitrary locations. We further show that the representation learned by NAFs can help improve visual learning with sparse views. Finally we show that a representation informative of scene structure emerges during the learning of NAFs. Project site: https://www.andrew.cmu.edu/user/afluo/Neural_Acoustic_Fields

1 Introduction

The sound of the ball leaving the bat, as much as its visible trajectory, tells us whether the hit is likely to be a home run or not. Our experience of the world around us is rich and multimodal, depending on integrated input from multiple sensory modalities. In particular, spatial acoustic cues provide us with a sense of the direction and distance of a sound source without needing visual confirmation, allow us to estimate the properties of a surrounding environment, and are critical to subjective realism in gaming and virtual simulations.

Recent progress in implicit neural representations has enabled the construction of continuous, differentiable representations of the visual world directly from raw image observations [1, 2, 3, 4]. However, our perception of the physical world is informed not only by our visual observations, but also by the spatial acoustic cues present in the environment. As a preliminary step in learning the acoustic properties of scenes, we explore an implicit model that represents the underlying impulse response of audio reverberations. As shown in Figure 1, our model can model the spatial propagation of sound in a physical scene.

Past work has explored capturing the underlying acoustics of a scene [5, 6, 7]. These models, however, require handcrafted parameterizations which, critically, prevent such approaches from being

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

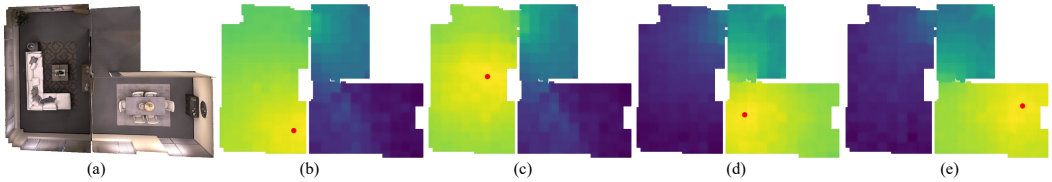

Figure 1: Neural Acoustic Field (NAF) learns an implicit representation for acoustic propagation. **(a)** A 3D top-down view of the house with two rooms. **(b)-(e)** The loudness of acoustic field as predicted by our NAF is visualized for an emitter located at the red dot. Notice how sound does not leak through walls, and the portaling effect open doorways can have. Louder regions are shown in yellow.

applied to arbitrary scenes. In this work, we extend this approach by constructing an implicit neural representation which captures, in a *generic manner*, the underlying acoustics of a scene.

Learning a representation of scene acoustics poses several challenges compared to the visual setting. First, how do we generate plausible audio impulse responses at each emitter-listener position? While we may represent the visual appearance of a scene with an underlying three-dimensional vector, an acoustic reverberation (represented as an impulse response) can consist of over 10,000 time-domain values and, thus, is significantly harder to capture. Second, how do we learn an acoustic neural representation that densely generalizes to novel emitter-listener locations? In the visual setting, ray-tracing can enforce view consistency across large portions of a visual scene (modulo occlusions). While in principle, in a similar manner, we may reflect acoustic "rays" in a scene represented as an implicit function to obtain an impulse response, an intractable amount of compute is necessary to obtain the desired representation [8].

To address both challenges, we propose Neural Acoustic Fields (NAFs). To capture the complex signal representation of impulse responses in a compact and spati ally continuous fashion, NAFs encode and represent an impulse-response in the time-frequency domain. Motivated by the strong influence of nearby geometry on anisotropic reflections [6], we propose to condition NAFs on local geometric information present at both the listener and emitter locations when decoding the impulse response. In our framework, local geometric information is learned directly from impulse responses. Such a decomposition facilitates the transfer of local information captured from training emitter-listener pairs to novel combinations of emitters and listeners.

We show that NAFs are able to outperform baselines in modeling scene acoustics, and provide detailed analysis of the design choices in NAFs. We further illustrate how the structure learned by NAFs can improve cross-modal generation of novel visual views of a scene. Finally, we illustrate how the learned representation of NAFs enable the downstream application of inferring scene structure.

2 Related Work

Audio Field Coding There is a rich history of sound field representation, encoding, and interpolation methods for 3D spatial audio. Some approaches seek to directly approximate the sound field [9, 10, 11] while adding handcrafted priors. Others adopt a parametric representation that seek to model only the perceptual cues [5, 6, 7, 12, 13]. Since the complete acoustic field of a scene is computationally prohibitive to simulate in real time, and expensive to store in full fidelity, these methods have typically relied on a handcrafted encoding of the acoustic field, prioritizing efficiency above reproduction fidelity. In recent years, there has been interest in using deep learning to directly learn a sound field from data, without making strong assumptions about the scene. However in practice, these approaches use either a stationary listener or emitter [14, 15]. In contrast, our work enables the querying of the sound field for arbitrary emitter and listener locations.

Implicit representations Our approach towards modeling the underlying acoustics a scene relies on the use of a neural implicit representations. Implicit representations have emerged as a promising representation of 3D geometry [16, 17, 18, 19, 20] and appearance [1, 2, 3, 4, 21] of a scene. Compared to traditional discrete representations, implicit representations are a continuous mapping capable of capturing data at an "infinite resolution". In [22] proposed a grid based representation for implicit scenes, while more recently [23] has adopted spatial conditioning for 3D image synthesis, where in both settings, the grid enables a higher-fidelity encoding of the scene. Our work also leverages local grids to model acoustics, but as an inductive bias and way to generalize to novel inputs.

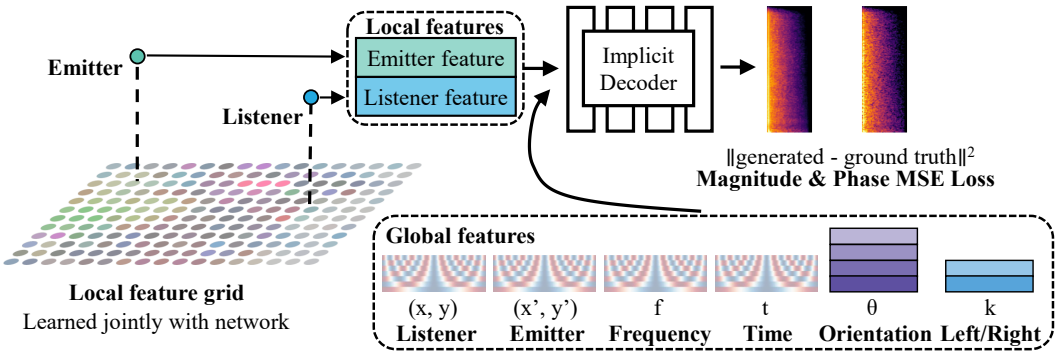

Figure 2: Overview of our NAF architecture where listener and emitter share a feature grid. Given a listener position and an emitter location, we first query a grid for local features which are learned together with the network during training. We compute the sinusoidal embedding of the positions, frequency, and time, and query a discrete embedding matrix using the orientation and left/right ear. Our method predicts magnitude and phase.

Audio-Visual Learning Our work is also closely related to joint modeling of vision and audio. By leveraging the correspondence between vision and audio, work has been done to learn unsupervised video and audio representations [24, 25], localize objects that emit sound [26, 27], and jointly use vision and audio for navigation [28]. Recent work aims to propose plausible reverberations or sounds from image input [29, 30], these approaches model the STFT using either convolution or implicit functions, which we also utilize. Different from them, our work leverages the geometric features learned by modeling acoustic fields to improve the learning of 3D view generation.

3 Methods

We are interested in learning a generic acoustic representation of an arbitrary scene, which can capture the underlying sound propagation of arbitrary sound sources across both seen and unseen locations in a scene. We first review relevant background information towards modeling environment reverberations. We then describe Neural Acoustic Fields (NAFs), a neural field which we show can capture, in a generic manner, the acoustics of arbitrary scenes. We further discuss how we can parameterize NAF so that it can capture acoustics property even at unseen sound sources and listener positions. Finally, we discuss the implementation details of our model illustrated in Figure 2.

3.1 Background on the Propagation of Sound

The sound emitted by a sound source undergoes decay, occlusion, and scattering due to both the geometric and material properties of a scene. For a fixed location pair $(\boldsymbol{q}, \boldsymbol{q}')$, we define the impulse-response at a listener position $\boldsymbol{q}$, as the sound pressure $p(t; \boldsymbol{q}, \boldsymbol{q}')$ induced by an impulse at $\boldsymbol{q}'$.

Given an accurate model of the impulse-response $p(t; \boldsymbol{q}, \boldsymbol{q}')$, we may model audio reverberation of any sound waveform $s(t)$ emitted at $\boldsymbol{q}'$, by computing the response $r(t, \boldsymbol{q}, \boldsymbol{q}')$ at listener location $\boldsymbol{q}$ by querying the continuous field and using temporal convolution:

$$r(t; \boldsymbol{q}, \boldsymbol{q}') = s(t) \circledast p(t; \boldsymbol{q}, \boldsymbol{q}') \tag{1}$$

3.2 Neural Acoustic Fields

We are interested in constructing a continuous representation of the underlying acoustics of a scene, which may specify the reverberation patterns of an arbitrary sound source. The parameterization of an impulse-response introduced in Section 3.1 provides us with a method to model audio propagation when given an omnidirectional listener and emitter. To construct a model of a directional listener, we need to further model the 3D head orientation $\theta \in \mathbb{R}^2$, and ear $k \in \{0, 1\}$ (binary left or right) of a listener, in addition to the spatial position $\boldsymbol{q} \in \mathbb{R}^3$ of the listener and $\boldsymbol{q}' \in \mathbb{R}^3$ of the emitter.

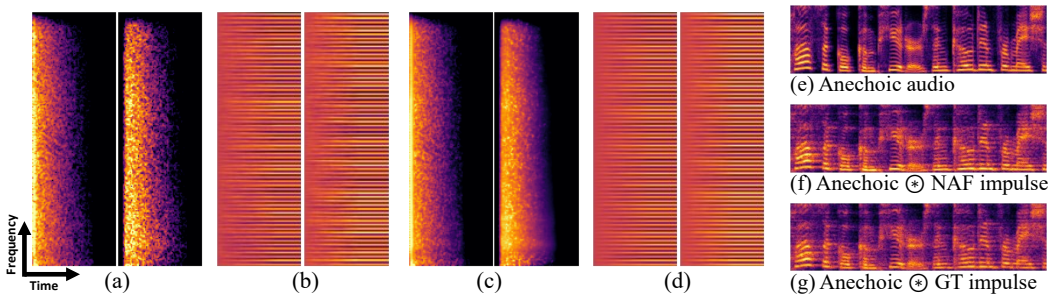

Figure 3: **Qualitative Visualization of Test Set Impulse Response Prediction.** **(a)** Ground truth log-magnitude. **(b)** Ground truth phase. **(c)** NAF predicted log-magnitude. **(d)** NAF predicted phase. **(e)** log-magnitude of an anechoic audio (without any reverberation). **(d)** The sound with a reverberation impulse response from our NAF. **(e)** The sound with the ground truth reverberation impulse response applied. Note phases are unwrapped for visualization purposes. Time & frequency are on the horizontal and vertical axes.

We may then model the time domain impulse response $\boldsymbol{v}$ using a neural field Φ which takes as input the listener and emitter parameters:

$$\Phi : \mathbb{R}^8 \times \{0, 1\} \rightarrow \mathbb{R}^T; (\boldsymbol{q}, \theta, k, \boldsymbol{q}') \rightarrow \Phi(\boldsymbol{q}, \theta, k, \boldsymbol{q}') = \boldsymbol{v} \tag{2}$$

Directly outputting the impulse-response waveform $\boldsymbol{v} \in \mathbb{R}^T$ in the time domain with a neural network is difficult due to its high dimensional (over 10,000 elements) and chaotic nature. A naïve solution would be to further add t as an additional argument to our neural field, but we found that such a solution worked poorly, due to the highly non-smooth representation of the waveform (see supplementary). We instead encode the impulse-response utilizing a short-time Fourier transform (STFT) denoted v_{STFT}, which we find to be significantly more amenable to neural network prediction due to the smoother nature of the time-frequency space. In Figure 3 we show magnitude spectrograms for ground truth impulse responses and those learned by our network. As v_{STFT} is a complex value, we further factorize v_{STFT} into log-magnitude and phase angle components. For phase angle, we use the instantaneous frequency (IF) representation proposed in GANSynth [31]. To compute the IF representation, the phase angle is unwrapped and has the finite difference taken across the time dimension for each frequency in the STFT. This transformation results in a phase representation that conducive to learning due to more regular structure.

Thus, our parameterization of NAF is a neural field Ω that is trained to estimate the impulse response function ϕ, and outputs $[\boldsymbol{v}_{\text{STFT_mag}}, \boldsymbol{v}_{\text{STFT_IF}}]$ for a given time and frequency coordinate:

$$\begin{aligned} &\Omega : \mathbb{R}^{10} \times \{0, 1\} \rightarrow \mathbb{C} \\ &(\boldsymbol{q}, \theta, k, \boldsymbol{q}', t, f) \rightarrow \Omega(\boldsymbol{q}, \theta, k, \boldsymbol{q}', t, f) \approx [\boldsymbol{v}_{\text{STFT_mag}}(t, f), \boldsymbol{v}_{\text{STFT_IF}}(t, f)] \end{aligned} \tag{3}$$

We train our model using MSE loss between the generated and ground truth spectrograms $\boldsymbol{v}_{\text{STFT}}$:

$$\begin{aligned} \mathcal{L}_{\text{NAF}} = &\|\Omega(\boldsymbol{q}, \theta, k, \boldsymbol{q}', t, f)_{\text{mag}} - \boldsymbol{v}_{\text{STFT_mag}}(t, f)\|^2 + \\ &\alpha \|\Omega(\boldsymbol{q}, \theta, k, \boldsymbol{q}', t, f)_{\text{IF}} - \boldsymbol{v}_{\text{STFT_IF}}(t, f)\|^2 \end{aligned} \tag{4}$$

across spectrogram coordinates t and f. Where α is a scaling value used to balance the two losses.

3.3 Generalization through Local Geometric Conditioning

We are interested in parameterizing the underlying acoustic field, so that we may not only accurately represent impulse-response at emitter-listener pairs we see during training, but also at novel combinations of emitter and listener seen at test time. Such generalization may be problematic when directly parameterizing NAFs utilizing a MLP with inputs specified in Eqn (3), as the network may learn to directly overfit and entangle the relation between emitter and listener impulse-responses.

What generic information may we extract from a given impulse-response between an emitter and listener? In principle, extracting the full dense geometric information in a scene would enable us to robustly generalize to new emitter and listener locations. However, the amount of geometric

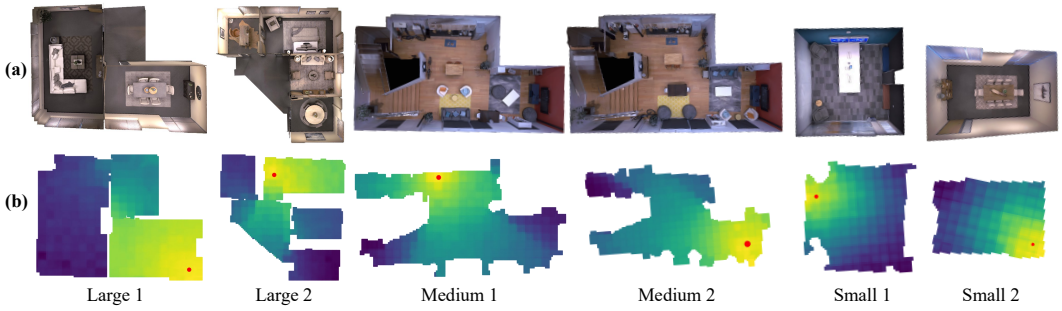

Figure 4: **Qualitative Visualization of Neural Acoustic Fields. (a)** Top down view of the rooms. **(b)** Results as inferred by our neural acooustic field. Loudness of a sound given a emitter location indicated in red, lighter color indicates louder sound. Note how openings and walls lead to portaling and occlusion of the sound.

information available in a particular impulse-response, especially for positions far away from either current emitter and listener is limited, since these positions have little impact on the underlying impulse-response. In contrast, the local geometry near either emitter and listener positions will have a strong influence in the impulse-response, as much of the anisotropic reflection comes from such geometry [32]. Inspired by this observation, we aim to capture and utilize local geometric information, near either emitter or listener locations, as a means to predict impulse-responses across novel combinations.

To parameterize and represent these local geometric features, we learn a 2D grid of spatial latents which we illustrate in Figure 2. The spatial latents are randomly initialized and uniformly distributed in the room. When predicting an impulse-response at a given emitter and offset position, we query the learned grid features at both emitter and listener positions, and provide it as additional context into our NAF network Ω. Such features provide rich information on the impulse-response, enabling NAF to generalize better to unseen combinations of both emitter and listener locations. In the rest of this work, we refer to the NAFs with local geometric features as Ω_{grid}. We learn grid latent features jointly with the underlying parameters of NAF. Additional details can be found in the supplementary.

Such a design choice, however, still requires us to consider how to further combine local geometric information captured separately from either listeners or emitters. A naïve implementation would be to maintain separate feature grids for both listener and emitter positions. Such an approach fails to account for the fact that the local geometric information captured by emitter may also inform the local geometric information around a listener. Examining Green's function, which is the solution to the wave equation, we note that it is in fact symmetric with respect to exchanging either listener or emitter positions [7], indicating that the impulse-response does not change when omnidirectional emitters and listeners are swapped (acoustic reciprocity). Such a result means that we may in fact utilize the local geometric information captured near an emitter position interchangeably for either emitters and listeners. Thus, we propose our local geometric information as a single latent grid, which we show to outperform the naïve dual grid implementation.

4 Experiments

In this section, we demonstrate that our model can faithfully represent the acoustic impulse response at seen and unseen locations. Additional ablation studies verify the importance of utilizing local geometric features to enable test time generation fidelity. Next, we demonstrate that learning acoustic fields could facilitate improved visual representations when training images are sparse. Finally we show that the learned NAF can be used to infer scene structure.

4.1 Setup

For each scene, we holdout 10% of the RIRs randomly as a test set. Each scene is trained for 200 epochs, which takes around 6 hours for the largest scenes on four `Nvidia V100s`. In each batch, we sample 20 impulse responses, and randomly select $2,000$ frequency & time pairs within each

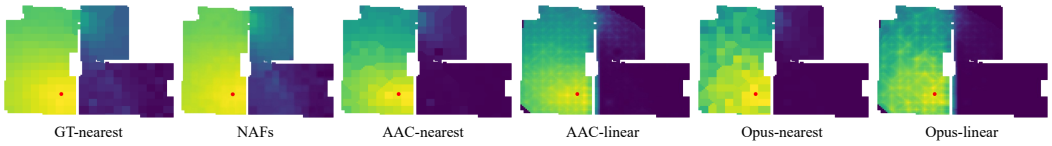

Figure 5: **Comparison of the acoustic fields.** From left to right, we visualize the loudness maps generated by the full ground truth, our NAFs, and by AAC or Opus coding combined with linear and nearest neighbor interpolation on the training set. Emitter location shown in red. Our method can faithfully reproduce the loudness map present in the ground truth.

spectrogram. An initial learning rate of 5×10^{-4} is used for the network and the grid features. We add a small amount of noise sampled from $\mathcal{N}(0, 0.1)$ to each coordinate during training to prevent degenerate solutions. For evaluating the learned acoustic fields, we use two different datasets:

Soundspaces. Soundspaces [28] is a synthetic dataset generated via ray-tracing, and is auralized with an ambisonic head-related transfer function (HRTF). This dataset consists of R_i probe points for each scene, with each probe capable of representing an emitter or listener location for up to R_i^2 emitter and listener pairs. The emitters are represented as omnidirectional, while the listener acts as a stereo receiver that can have one of four different orientations. The listeners and emitters are at fixed height. Our NAFs are trained on 6 representative scenes, where 2 consist of multi-room layouts; 2 consist of a single room with a non-rectangular walls; and 2 consist of a single room with rectangular walls as in Figure 4.

MeshRIR. The MeshRIR dataset [33] is recorded from a real scene, and contains monaural data collected from a cuboidal room. The listener locations are at fixed height. The emitters surround the listeners both above and below the listener plane.

4.2 Architecture Details

The Soundspaces dataset lacks the full parameterization of an acoustic field described in Equation 3, so we train NAF with a restricted parameterization that is available in the dataset. This allows for two degrees of freedom along the $x - y$ plane for the listener locations $q \in \mathbb{R}^2$ and the emitter location $q' \in \mathbb{R}^2$. The listener is binaural with $k \in \{0, 1\}$, and can assume four possible orientations $\theta \in \{0, 90, 180, 270\}$, while the emitter is omnidirectional. In particular, we utilize a parameterization of Ω_{grid} which maps an input tuple $[x, y, x', y', f, t] \times \{0, 90, 180, 270\} \times \{0, 1\}$ to two scalar values that represents the magnitude and phase for a given time and frequency in the STFT:

$$\Omega_{\text{grid}}(x, y, \theta, k, x', y', t, f) \Rightarrow [\boldsymbol{v}_{\text{STFT_mag}}(t, f), \boldsymbol{v}_{\text{STFT_IF}}(t, f)] \tag{5}$$

To encode the rotation θ, as there are only 4 possible discrete rotations in the dataset, we directly query into a learnable embedding matrix of shape $\mathbb{R}^{4 \times n}$, returning a $\mathbb{R}^{1 \times n}$ vector. Similarily, to encode the left and right ear k, we similarly query into a learnable embedding matrix of shape $\mathbb{R}^{2 \times n}$, returning a $\mathbb{R}^{1 \times n}$ vector. The f, t tuple representing the frequency and time respectively are scaled to $(-1, 1)$ and processed with sinusoidal encoding using 10 frequencies of $\sin$ and $\cos$. For MeshRIR, we set emitter $q' \in \mathbb{R}^3$ to account for the emitters that can vary in height, and do not utilize the orientation or binaural embedding.

To obtain local geometric features for either an emitter or listener in a scene, we assume that our scene is contained within a set of pixels $\mathcal{P} = \{P_1 ... P_k\}$ which form a grid over the scene. For a given position tuple (x, y) as query location, we then interpolate the local features. Where $\mathcal{L}(\cdot)$ is the interpolation function. $(p_1^* \dots p_k^*)$ are the set of all pixel that form the grid, and $\tilde{f}(\cdot)$ represents the features stored at a given pixel:

$$(x, y) \Rightarrow \mathcal{L}(x, y; \tilde{f}(p_1^*), \dots \tilde{f}(p_k^*)) \tag{6}$$

$$= \sum_{i=1}^{k} w_i \tilde{f}(p_i^*) \tag{7}$$

	Large 1		Large 2		Medium 1		Medium 2		Small 1		Small 2		MeshRIR		Mean	
Model	Spec.↓	T60↓	Spec.↓	T60↓	Spec.↓	T60↓	Spec.↓	T60↓	Spec.↓	T60↓	Spec.↓	T60↓	Spec.↓	T60↓	Spec.↓	T60↓
AAC-nearest	1.913	9.996	1.989	13.31	2.111	6.148	2.122	6.051	2.296	9.798	2.509	5.809	1.057	4.740	1.999	7.979
AAC-linear	1.904	8.847	1.964	11.63	2.105	4.585	2.116	4.422	2.299	8.253	2.521	6.021	1.081	6.697	1.998	7.208
Opus-nearest	1.740	12.20	1.817	15.15	1.887	7.875	1.898	7.897	2.058	10.68	2.238	7.564	1.711	5.068	1.907	9.493
Opus-linear	1.780	11.30	1.827	13.55	1.922	6.710	1.934	6.917	2.097	9.116	2.284	6.981	1.743	5.768	1.941	8.621
DSP	1.106	14.62	1.170	13.68	1.064	10.24	1.067	9.732	1.079	12.77	1.097	11.03	N/A	N/A	1.097	12.01
NAF (Dual)	0.413	6.288	0.421	7.111	0.386	3.173	0.387	3.169	0.365	3.497	0.361	2.210	**0.403**	4.201	0.388	4.241
NAF (Shared)	**0.396**	**4.166**	**0.413**	**6.075**	**0.384**	**3.110**	**0.384**	**3.072**	**0.356**	**3.378**	**0.344**	**2.098**	**0.403**	**4.191**	**0.380**	**3.650**

Table 1: **Quantitative Results on Test Set Accuracy.** We report the spectral loss between generated and ground truth log spectrograms across methods, as well as the percentage (%) difference for the T60 reverberation time. The best method for each room is **bolded**. For the nearest and linear baselines, we perform interpolation in the time domain using samples from the training set. The DSP is not implemented for MeshRIR due to the lack of absolute room coordinates.

w_i is determined by a Nadaraya-Watson estimator with a Gaussian weighting kernel applied to the distance between query and grid coordinates:

$$w_i = K((x, y), (x_i, y_i)) / \sum_{j=1}^{k} K((x, y), (x_j, y_j)) \tag{8}$$

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 / 2\sigma^2) \tag{9}$$

Because this interpolation function is differentiable, we jointly learn the grid features during training. These queried features are combined with the coordinates processed with sinusoidal encoding using 10 frequencies of $\sin$ and $\cos$ functions. We process both the listener and emitter position tuples this way. We combine the grid based features with the sinusoidal embeddings and the discrete indexed embeddings as the input to our multilayer perceptron f_ϕ. Please refer to Figure 2 for a visualization of our model, and supplementary for further details. We compare using a shared local geometric feature with the emitter and listener, as well as having the emitter and listener query their own individual grids.

4.3 Evaluation of Neural Acoustic Fields

We first validate that we can capture environmental acoustics at unseen emitter-listener positions.

Baselines. We compare our model against two widely used high performance audio coding methods: Advanced Audio Coding (AAC) and Xiph Opus. We use low bitrates in order to attempt to approach the storage costs for our NAF. For each method, we apply both linear and nearest neighbor interpolation to the coded acoustic fields. Both linear and nearest neighbor approaches are widely used [34, 35, 36] in modeling of spatial audio. We further implement the binaural DSP baseline described by [14], which uses the image source method and a KEMAR HRTF. We also compare the listener and emitter either sharing or using individual local geometric features in our NAFs.

Each method is provided with the same train-test split. We visualize the acoustic fields produced by different methods in Figure 5. Details of our baselines can be found in the supplementary section G.

Metrics. We evaluate the results of our synthesis by measuring the spectral loss [37] between the generated and the ground truth log-spectrograms, as well as measuring the percentage error between the T60 reverberation time in the time domain. In this case, lower spectral loss and T60-error values indicates a better result. Additional quantitative results can be found in the supplementary. We also perform a human evaluation where subjects are presented with a two-alternative forced-choice task. Each trial requires selecting if the NAFs or the AAC-nearest auralized music samples best match with ground truth auralization.

Results. As shown in Table 1, our NAFs achieve significantly higher quality on the modeling of unseen impulse responses compared to strong interpolation baselines across all six scenes. Observing the qualitative results in Figure 4, we observe that NAFs can predict smoothly varying acoustic fields that are affected by the physical surroundings. Extending our model to MeshRIR which is captured from a real scene, we observe that our NAFs

Method	Storage (MiB)
AAC	312.07
Opus	163.23
NAF (Shared)	8.41

Table 3: **Average space consumption across 6 Soundspaces scenes. Lower is better.**

	Large Room 1						Large Room 2					
	PSNR ↑			SSIM ↑			PSNR ↑			SSIM ↑		
Training Images	75	100	150	75	100	150	75	100	150	75	100	150
NeRF	25.41	27.36	29.85	0.872	0.892	0.926	25.70	27.74	29.34	0.821	0.853	0.879
NeRF + NAF	**26.19**	**27.59**	**29.90**	**0.895**	**0.911**	**0.927**	**26.24**	**28.22**	**29.45**	**0.837**	**0.866**	**0.879**

Table 2: **Quantitative Results on Cross-Modal Image Learning.** Quantitative results on joint training of NeRF and NAF jointly conditioned on a single local grid. We use very sparse training images in highly complex scenes. When evaluated on 50 test images, we observe that cross-modal learning helps improve PSNR and SSIM when the visual training data is more sparse.

continue to perform better on both spectral and T60 metrics. Comparing our results against baselines in Figure 5 and Table 3, our methods are able to better approximate the ground truth at a fraction of the storage cost, and does not exhibit the sound energy leakage present in linear interpolation. The size of a spatial acoustic field is important for real life applications. In our human evaluation with 21 subjects who were asked to judge 10 test-time RIRs, 82.38% of responses indicates that our NAFs were higher quality compared to AAC-Nearest.

A comparison of using shared and dual local geometric features indicates that despite having fewer learnable parameters, we achieve better performance by sharing the local geometric features. Examples of individual impulse responses generated by our model are shown in Figure 3.

Generalization through Geometric Conditioning. We next assess the impact of utilizing local geometric conditioning as a means to generalize to novel combinations of emitter-listener positions. On the "Large 1" room, in Figure 7 we evaluate test set spectral error when NAF is trained with a limited percentage of the training data either with or without local geometric conditioning. We find that such geometric conditioning enables better test set reconstruction error, with the performance gap increasing with less data.

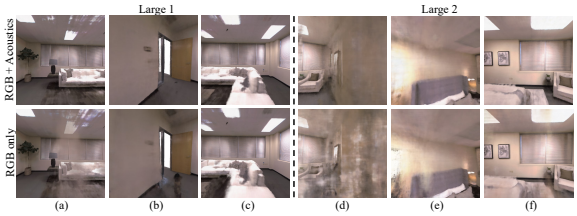

Figure 6: **Qualitative Visualization of Cross-Modal Image Learning.** Qualitative comparison between NeRF+NAFs with RGB and acoustic supervision, and NeRF learned with only RGB supervision. **(a)-(c)** Three views from "Large 1". **(d)-(f)** Three views from "Large 2".

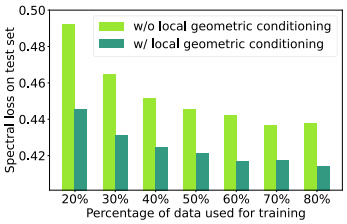

Figure 7: **Local Geometric Conditioning.** Comparison of NAF with and without local geometric conditioning trained with different amounts of data.

4.4 Cross-modal learning

In this experiment, we explore the effect of jointly learning acoustics and visual information when we are given sparse visual information. Recall that our NAF includes a local geometric feature grid $\mathcal{P}$ that covers the entire scene. For our cross-modal learning experiment, we jointly learn this feature grid with a NeRF network modified to accept local features sampled from this grid along with the traditional sinusoidal embedding. In the acoustics branch, we query the grid using emitter and listener positions. In the NeRF branch, we use point samples along the ray projected on the grid plane to query the features. In both cases, the process is fully differentiable. We use a standard implementation of NeRF with a coarse and fine network.

In the NeRF only setting, we minimize color C reconstruction loss for a ray r over a batch of rays $\mathcal{R}$: $\mathcal{L}_{\text{RGB}} = \sum_{r \in \mathcal{R}} ||\hat{C}(r) - C(r)||_2^2$. In contrast, in the NAF + NeRF experiment, we jointly minimize $\mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{NAF}}$, where $\mathcal{L}_{\text{NAF}}$ is defined in equation 4. We utilize 64 coarse samples and 128 fine samples for each ray, and sample 1024 rays per batch.

Results. We train on the two large rooms in our training set. For each room $75, 100, 150$ images are used for training, while the same 50 images of novel views are used for all configurations during

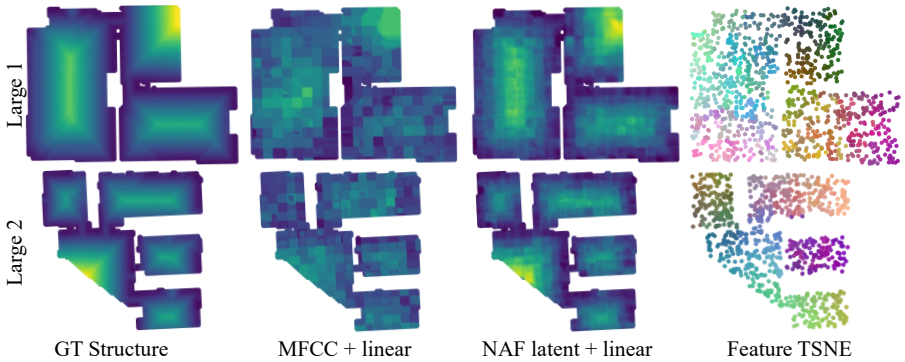

Figure 8: **Visualization of scene structure decoding with a linear layer. Column 1:** The ground truth scene structure map, at each position we visualize the distance to the nearest wall. **2:** Linear decoding results using MFCC features. **3:** Linear decoding results using NAF features. **4:** TSNE applied to the NAF features.

testing. In Table 2 we observe that training with acoustic information helps improve the PSNR and SSIM of the visual output. This effect is more significant when the training images are very sparse, the NAF network helps less when there is sufficient visual information. Qualitative results are shown in Figure 6, we see there is a reduction of floaters in free space.

4.5 Inferring scene structure

Given a reverberant sound, humans are able to build a mental representation of the surrounding room and make a judgement about the distance of nearby obstacles [38]. We investigate the intermediate representations constructed by our neural network in the process of learning an acoustic field, and examine if these representations can be used to decode the scene structure.

Setup. The intermediate representation of the NAF depends on both listener locations q and emitter location q', the rotation angle θ, the ear k, the time t and frequency f. For consistency, at a given location (x^*, y^*) in the scene, we extract the NAF latent by setting the emitter location $q'_i = (x^*, y^*)_i$. For the listener location, we iterate over five randomly selected points in the scene $q \in [q_1, \ldots, q_5]$, which we keep constant for all q'_i. The rotation angle is fixed to $\theta = 0$, and we compute the representation average over all possible (k, t, f), and concatenate latents for the selected q. For our NAFs, latents are extracted from the last layer prior to the output which includes 512 neurons. As a

	Explained variance						
Features	Large 1	Large 2	Medium 1	Medium 2	Small 1	Small 2	Mean
MFCC	0.501	0.458	0.614	0.642	0.820	0.723	0.626
NAF latents	**0.908**	**0.891**	**0.900**	**0.923**	**0.936**	**0.916**	**0.913**

Table 4: **Quantitative Results on scene structure decoding.** We measure the explained variance scores of the predicted wall distance against the ground truth wall distance at test time locations after linear decoding. NAF latents consistently achieve higher explained variance scores than MFCC features.

comparison to our learned representation, we extract Mel-frequency cepstral coefficients (MFCCs) from the ground truth impulse response provided by a nearest neighbor interpolator. We use a similar setup as above, for a given location we set this to be q'_i, and iterate over the same five listener locations $q_{1\ldots5}$. We average the MFCCs over the left and right ear, and concatenate for the selected q. After flattening, the MFCC features are approximately 500 dimensional for any given room.

We fit a single linear layer to NAF and MFCC features respectively. For testing and visualization of the linear decoding results, we sample a regular grid of points with $0.1m$ distance between each point. For fitting the linear decoder, we randomly sample points within the scene such that the number of training points are 10% as many as the testing points. For each location in the scene, we extract the distance to the nearest wall as the decoding target.

Results. We visualize the results of our linear decoding in Figure 8. The intermediate representation of our NAFs reveals an underlying structure that is both smooth and semantically meaningful. In the multiroom scenes, the latent is well separated for each room. We are able to successfully decode the scene structure with a linear layer when using our NAFs, but decoding fails when using MFCC features. In Table 4, we show the amount of explained variance of our decoding results on the test set. Our learned features are able to consistently achieve much higher scores than those using MFCC features.

5 Discussion

Limitations and Future Work. Although our method achieves generalization and high quality representations of acoustic fields within a single scene, NAFs do not currently generalize to multiple scenes. Future work may explore generalization to novel scenes. One possible approaches may be to incorporate multi-modal inputs with the goal of synthesizing an acoustic field with few-shot visual or acoustic input.

Societal impact. Our work focuses on learning a high quality representation of acoustic fields. The primary use case for our work lies in virtual reality and gaming. As our work can lead to more believable and higher quality representations of spatial audio than alternative methods, it is possible that our work could increase the dependency and time spent on gaming. The more compact nature of our acoustic representations may allow for spatial audio to be deployed to more systems, and enable more equitable access.

Conclusion. In summary, this paper introduces Neural Acoustic Fields (NAFs), a compact, continuous, and differentiable acoustic representation which can represent the underlying reverberation of different audio sources in a scene. By conditioning NAFs locally on the underlying scene geometry, we demonstrate that our approach enables the prediction of plausible environmental reverberations even at unseen locations in the scene. Furthermore, we demonstrate that the acoustic representations learned through NAFs are powerful, and may be utilized to facilitate audio-visual cross-modal learning, as well as to infer the structure of scenes.

Acknowledgments and Disclosure of Funding

We would like to thank Leila Wehbe for providing feedback on our paper. This work was supported by the MIT-IBM Watson AI Lab, Amazon Research Award, ONR MURI, DARPA Machine Common Sense program, ONR (N00014-18-1-2847), NIDCD R01 DC015988, Mitsubishi Electric, and the Tianqiao and Chrissy Chen Institute.

References

[1] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Proc. NeurIPS 2019*, 2019. 1, 2

[2] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020. 1, 2

[3] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. CVPR*, 2020. 1, 2

[4] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Proc. NeurIPS*, 2020. 1, 2

[5] Nikunj Raghuvanshi and John Snyder. Parametric wave field coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 1, 2

[6] Nikunj Raghuvanshi and John Snyder. Parametric directional coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 1, 2

[7] Chakravarty R Alla Chaitanya, Nikunj Raghuvanshi, Keith W Godin, Zechen Zhang, Derek Nowrouzezahrai, and John M Snyder. Directional sources and listeners in interactive sound propagation using reciprocal wave field coding. *ACM Transactions on Graphics (TOG)*, 39(4):44–1, 2020. 1, 2, 5

[8] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proc. CVPR*, 2021. 2

[9] Rémi Mignot, Gilles Chardon, and Laurent Daudet. Low frequency interpolation of room impulse responses using compressed sensing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):205–216, 2013. 2

[10] Niccolo Antonello, Enzo De Sena, Marc Moonen, Patrick A Naylor, and Toon Van Waterschoot. Room impulse response interpolation using a sparse spatio-temporal representation of the sound field. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1929–1941, 2017. 2

[11] Natsuki Ueno, Shoichi Koyama, and Hiroshi Saruwatari. Kernel ridge regression with constraint of helmholtz equation for sound field interpolation. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–440. IEEE, 2018. 2

[12] Ravish Mehra, Lakulish Antani, Sujeong Kim, and Dinesh Manocha. Source and listener directivity for interactive wave-based sound propagation. *IEEE transactions on visualization and computer graphics*, 20(4):495–503, 2014. 2

[13] Anton Ratnarajah, Shi-Xiong Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. Fast-rir: Fast neural diffuse room impulse response generator. *arXiv preprint arXiv:2110.04057*, 2021. 2

[14] Alexander Richard, Dejan Markovic, Israel D Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando Torre, and Yaser Sheikh. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*, 2020. 2, 7

[15] Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. Deep impulse responses: Estimating and parameterizing filters with deep networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3209–3213. IEEE, 2022. 2

[16] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proc. ICCV*, 2019. 2

[17] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proc. CVPR*, pages 5939–5948, 2019. 2

[18] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. CVPR*, 2019. 2

[19] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. ICCV*, pages 2304–2314, 2019. 2

[20] Yining Hong, Yilun Du, Chunru Lin, Joshua B Tenenbaum, and Chuang Gan. 3d concept grounding on neural fields. *arXiv preprint arXiv:2207.06403*, 2022. 2

[21] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2

[22] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser. Local implicit grid representations for 3d scenes. In *Proc. CVPR*, pages 6001–6010, 2020. 2

[23] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W Taylor, and Joshua M Susskind. Unconstrained scene generation with locally conditioned radiance fields. *arXiv preprint arXiv:2104.00670*, 2021. 2

[24] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29:892–900, 2016. 3

[25] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 3

[26] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 3

[27] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 3

[28] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020. 3, 6

[29] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-model reverb impulse response synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 3

[30] Yilun Du, M. Katherine Collins, B. Joshua Tenenbaum, and Vincent Sitzmann. Learning signal-agnostic manifolds of neural fields. In *Advances in Neural Information Processing Systems*, 2021. 3

[31] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*, 2019. 4

[32] Juhani Paasonen, Aleksandr Karapetyan, Jan Plogsties, and Ville Pulkki. Proximity of surfaces—acoustic and perceptual effects. *Journal of the Audio Engineering Society*, 65(12):997–1004, 2017. 5

[33] Shoichi Koyama, Tomoya Nishida, Keisuke Kimura, Takumi Abe, Natsuki Ueno, and Jesper Brunnström. Meshrir: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE, 2021. 6

[34] Lauri Savioja, Jyri Huopaniemi, Tapio Lokki, and Ritta Väänänen. Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47(9):675–705, 1999. 7

[35] Nikunj Raghuvanshi, John Snyder, Ravish Mehra, Ming Lin, and Naga Govindaraju. Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes. In *ACM SIGGRAPH 2010 papers*, pages 1–11. 2010. 7

[36] Christoph Pörschmann, Johannes M Arend, David Bau, and Tim Lübeck. Comparison of spherical harmonics and nearest-neighbor based interpolation of head-related transfer functions. In *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2020. 7

[37] Alexandre Défossez, Neil Zeghidour, Nicolas Usunier, Léon Bottou, and Francis Bach. Sing: Symbol-to-instrument neural generator. *arXiv preprint arXiv:1810.09785*, 2018. 7

[38] Andrew J Kolarik, Brian CJ Moore, Pavel Zahorik, Silvia Cirstea, and Shahina Pardhan. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics*, 78(2):373–395, 2016. 9

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See repo provided on site.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] .
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We use a single fixed seed for all our experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] We utilize public datasets only.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Does not apply to our data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] Was not determined to create any risks.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No]