ActiveRIR: Active Audio-Visual Exploration for Acoustic Environment Modeling

Arjun Somayazulu[1], Sagnik Majumder[1,2], Changan Chen[1], Kristen Grauman[1,2]

***Abstract*— An *environment acoustic model* represents how sound is transformed by the physical characteristics of an indoor environment, for any given source/receiver location. Traditional methods for constructing acoustic models involve expensive and time-consuming collection of large quantities of acoustic data at dense spatial locations in the space, or rely on privileged knowledge of scene geometry to intelligently select acoustic data sampling locations. We propose *active acoustic sampling*, a new task for efficiently building an environment acoustic model of an unmapped environment in which a mobile agent equipped with visual and acoustic sensors jointly constructs the environment acoustic model and the occupancy map on-the-fly. We introduce ActiveRIR, a reinforcement learning (RL) policy that leverages information from audio-visual sensor streams to guide agent navigation and determine optimal acoustic data sampling positions, yielding a high quality acoustic model of the environment from a minimal set of acoustic samples. We train our policy with a novel RL reward based on information gain in the environment acoustic model. Evaluating on diverse unseen indoor environments from a state-of-the-art acoustic simulation platform, ActiveRIR outperforms an array of methods—both traditional navigation agents based on spatial novelty and visual exploration as well as existing state-of-the-art methods.**

I. INTRODUCTION

The acoustic properties of the sounds a mobile agent hears are determined by the physical characteristics of the space it is in, such as the room's geometry, the objects within it, the types of materials that comprise the room and objects' surfaces, and the agent's proximity and orientation with respect to the sound source. Consider a conversation with someone standing across a large auditorium or gym, compared to the same conversation sitting a few feet apart in a small, carpeted living room; Large spaces with hard surfaces (*e.g.* concrete, glass) will add reverberation to audio, while small, cluttered spaces covered in soft materials (*e.g.* curtains, carpet) absorb sound waves quicker, producing dull and anechoic audio. These physical properties transform any sound emitted from a source location in the environment according to various acoustic phenomena, including direct sounds, early reflections, and late reverberations, before it reaches our ears. Together, these phenomena form a Room Impulse Response (RIR), which characterizes the transfer function between a sound emitted at the source location and the sound that reaches a microphone or our ears [1].

[1]The University of Texas at Austin
[2]FAIR, Meta

An *environment acoustic model* is a complete representation of a scene's acoustics [2], [3], [4], [5], [6]. Given a sound source location and a listener's position and orientation (pose) as a query, the model renders the corresponding RIR, accounting for all major acoustic phenomena due to the physical properties of the space.

Environment acoustic models are critical for robotics. In mobile robotics, an environment acoustic model can provide rich contextual information to an agent. Tasks such as navigating to a target sound [7] and separating out a sound of interest from background sounds [8], [9] require an agent to decide where to move based on the audio it hears. An agent equipped with the environment acoustic model can better anticipate the effects of its movement on the observed audio. In AR/MR applications, acoustic models allow a virtual sound source (such as a human speaker) inserted at a position in the user's real-world space to sound properly spatialized with acoustics that match the space, as the user moves around their environment.

Capturing an RIR is a physically involved process. A loudspeaker must be placed at the desired height and location, and a microphone setup placed at a similar height at the desired receiver location and orientation. The speaker emits a sound impulse, and the receiver microphone(s) record the resulting RIR, which can last several seconds.

Existing methods for environment acoustic modeling assume extensive physical access to the environment in order to collect these RIRs at arbitrary positions. Neural Field methods [3], [10] require large quantities of RIRs captured at dense source/listener locations (approx. every 1 meter) throughout the environment, which can be expensive and time-consuming. A few-shot approach [2] overcomes the intensive data requirement, building acoustic models of novel environments from a limited number of observations, but still requires extensive knowledge of the floorplan in order to sample uniformly spaced locations around obstacles. Other methods rely heavily on knowledge of the scene geometry from 3D meshes [4] or floorplans [6]. Importantly, these methods all assume *instantaneous access to observations at arbitrary positions in the environment*, which is unrealistic in the embodied robotics context—where an agent must physically travel between locations—and in unmapped environments where no prior knowledge of the floorplan and obstacles is available.

We introduce *active acoustic sampling*, a new task that

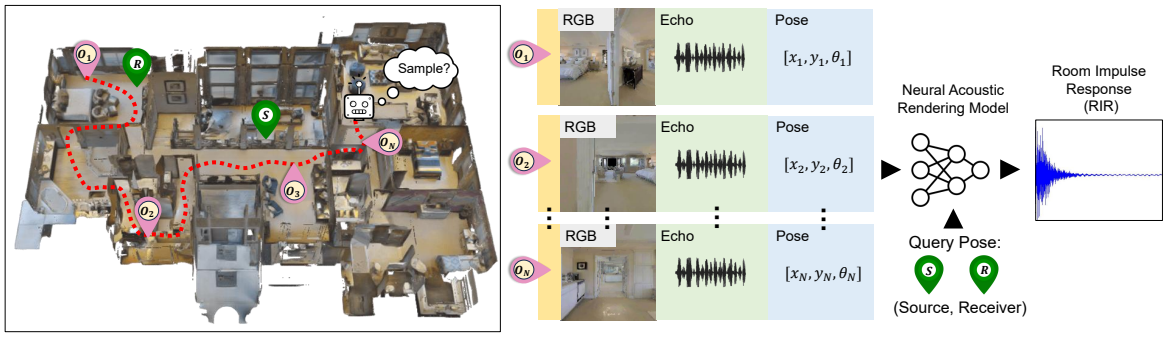

Fig. 1. **Active acoustic sampling.** An agent must intelligently navigate an unmapped 3D scene and actively sample audio-visual observations (the scene's *acoustic context*) to construct an acoustic model of the environment, given limited navigation time and a fixed sampling budget. When queried with an arbitrary sound source position and receiver pose in the space, the learned environment acoustic model should accurately generate the corresponding Room Impulse Response (RIR) at that pose.

requires a single mobile agent with audio-visual sensing to efficiently construct an unmapped environment's acoustic model within a total budget of acoustic samples, despite only on-the-fly discovery of its floorplan. The proposed problem is distinct from traditional visual exploration [11], [12], [13], where agents prioritize motions in a scene to rapidly complete the occupancy map. While in floorplan mapping the best places to reach are those that add visibility to the widest floor area, in acoustic modeling the most valuable poses (sampling spots) in the environment depend on all aspects of the 3D geometry and surface materials.

To address this challenge, we present ActiveRIR, an active sampling policy that can be deployed on mobile agents in environments that are both *unseen* and *unmapped*. ActiveRIR is trained with a novel audio-visual exploration reward to guide wide agent exploration and inform the agent's decision on when to sample an acoustic observation, within a budget of total acoustic samples. Our acoustic reward measures the global improvement in the agent's environment acoustic model estimate after an observation is sampled, ensuring that the limited context used to build the environment acoustic model contains only the most valuable observations seen by the agent during its exploration.

Evaluating on a diverse set of unseen and unmapped scanned real-world 3D indoor environments together with state-of-the-art (SOTA) acoustic [14] and visual [15] scene simulation platforms, ActiveRIR produces a higher-quality acoustic model in >70% fewer steps than passive approaches, and outperforms both traditional visual and spatial exploration-based methods [16], [11], [13], [12] as well as SOTA scene acoustic modeling methods [2]. We also demonstrate that the performance gain achieved using ActiveRIR-collected observations generalizes across multiple acoustic rendering methods, showing promising potential for ActiveRIR to be plugged as a module into existing acoustic rendering methods and improve the quality of generated environment acoustic models.

II. RELATED WORK

A. Environment acoustics and mapping

Estimating room acoustics from images of indoor scenes has been widely explored, including explicit RIR estimation from images [17], [18] as well as methods that model its acoustic transformation on source sounds or speech [19], [20], [21], [22], [23]. While these methods can produce audio that perceptually matches the general acoustics of the space (*e.g.* a concert hall vs. bedroom), they cannot reason about fine-grained acoustic effects of shifts in speaker or listener pose *within* a visual scene.

Audio field coding approaches [24], [25], [26] estimate generic perceptual RIR features using parametric sound field representations that model spatial relationships between acoustics at different positions. Recent approaches use Neural Fields to generate RIRs [3] directly, though they still require large-scale acoustic data (>10k samples) from dense spatial locations, and the learned model cannot generalize to other environments.

A transformer-based approach produces an acoustic model of an unseen environment given limited acoustic data sampled randomly from the floorplan [2]. Other GNN-based approaches generate RIRs given a full 3D scene mesh and dataset of acoustic material coefficients [4], [5]. These methods rely on prior knowledge of scene geometry, and assume the ability to instantaneously access visual and acoustic observations at selected positions in the space. Access to this privileged information—as well as the ability to teleport to new locations without penalty—are significant assumptions that are unrealistic for robots. In principle, the process that produced a pre-computed floorplan map or mesh could have also pre-computed the acoustic model with densely sampled observations using minimal additional time or energy. In short, assuming access to full scene geometry but *not* complete RIRs as well fails to capture real-world constraints in robotics.

B. Audio-visual navigation and exploration

Equipping mobile agents with sound production and audio capture has led to a proliferation in audio-visual embodied tasks, such as audio-visual sound source separation [9], [8], audio-visual navigation [27], [7], [28], [29] and audio-visual floorplan mapping [30], [31]. A mobile audio-visual agent can decide where to emit and receive acoustic observations to help with floorplan mapping [32]. While mapping the floorplan requires dense exploration of the space and its frontiers, we hypothesize that an accurate environment acoustic model of the same space can be built with far fewer acoustic observations sampled intelligently at select locations. Prior work [33] trains a policy to construct an environment acoustic model that relies on two mobile agents: an "emitter" emits an impulse sound and the "receiver" records the RIR. The agents navigate according to a local acoustic reward that measures the predicted RIR error at the agents' next position. However, unlike our global formulation, movement to minimize local acoustic prediction accuracy often prevents the agent from navigating towards areas with challenging and dynamic acoustics (*e.g.* hallways with turns and corners), hurting performance (see Sec. V).

III. Active acoustic sampling task

We introduce the task of *active acoustic sampling*. The goal of this task is to train a mobile agent to navigate an unmapped 3D environment (such as a home or office) and intelligently sample egocentric audio-visual observations ("acoustic context") within a predefined *sample budget* and *time budget*, such that a acoustic rendering model conditioned on this context can produce accurate and high-fidelity RIRs given arbitrary query source/receiver poses. Each audio-visual observation consists of the agent's egocentric RGB-D image and an echo response, namely an RIR obtained by placing the sound source and receiver microphones at the location of image capture, emitting a sinusoidal frequency sweep signal and recording the echoes [2]. See Fig. 1. Our sample budget is much lower than the timestep budget for navigation, requiring the agent to carry out the memory-intensive [33] task of capturing and storing echo responses only at a select few locations most valuable for the global scene acoustic model. The audio sampling is preemptive, *i.e.* the model decides to choose or skip an audio sample before capturing it.

Formally, given navigation time budget T and audio sample budget $N \ll T$, the agent must navigate an unmapped scene and collect audio-visual samples $\mathcal{C} = \{C_i\}^N$, where $C_i = (A_i, V_i, P_i)$. A_i denotes the binaural (two-channel) echo response, and V_i is the 90° FoV RGB-D image captured at the camera pose $P_i = (x_i, y_i, \theta_i)$ at location (x_i, y_i) and orientation θ_i.

Using $\mathcal{C}$ as the context, the agent must infer the scene's environment acoustic model, such that it can accurately estimate the RIR R^Q for arbitrary query $Q = (s_j, l_k)$, where $s_j = (x_j, y_j)$ is an omni-directional sound source at location (x_j, y_j), and $l_k = (x_k, y_k, \theta_k)$ is the receiver microphone pose. Thus, the task objective is to learn a policy π that guides our agent towards acoustically informative locations in the scene and helps it decide where to sample audio-visual observations to add to the acoustic context $\mathcal{C}$, such that a acoustic rendering model f conditioned on $\mathcal{C}$ approximates the ground-truth RIR R^Q for an arbitrary query Q, or $\tilde{R}^Q = f(Q|\mathcal{C})$. At each time step t, where $1 \leq t \leq T$, the policy can take action $\alpha_t \in \mathcal{A}$, where $\mathcal{A} = \{MoveForward, TurnLeft, TurnRight\} \times \{Sample, Skip\}$ is the agent's action space.

This task relies on context cues from both audio and vision. RGB-D images capture local room geometry, the presence of furniture/obstacles, and the materials of surfaces in view, while acoustic observations help the agent associate these physical properties with their acoustic effects on an emitted sound. Audio also captures longer-range room geometry beyond the agent's current FoV, which vision cannot. Exploiting this visual-acoustic correspondence is key in helping the agent decide how to move and where to *Sample* an acoustic observation.

IV. Approach

We pose the task as a reinforcement learning problem, where a model sequentially decides where to sample audio-visual observations given visual RGB-D frames and previously sampled audio[1], and the environment acoustic model is built from this collected context. We propose **ActiveRIR**, which consists of **1)** an audio-visual sampling policy π, and **2)** a neural acoustic rendering model f for predicting scene acoustics using the samples. See Fig. 2. Next, we describe these components in detail.

A. Audio-visual sampling policy

1) Policy inputs and architecture: At every time step t, our audio-visual sampling policy π receives O_t as the input and decides how to move in the environment, and whether to capture and add the echo response at the current step to our acoustic context (i.e., *Sample* or *Skip*). Formally, $O_t = (A_{t-1}, V_t, P_t)$, if the echo response was sampled at the previous step, and $O_t = (V_t, P_t)$ if it was skipped. Before encoding O_t, π pre-processes V_t to generate $V_t^\pi = (V_t^\mathcal{R}, V_t^M)$, where $V_t^\mathcal{R}$ is the RGB image from V_t, and V_t^M is a topdown global occupancy[2] map. Note that the map begins empty and accumulates as the agent selects its motions.

First, we feed A_{t-1} (if sampled), V_t^π and P_t to separate encoders and extract audio features a_{t-1}, visual

[1]All audio inputs to our model are represented as two-channel magnitude spectrograms computed using the short-time Fourier transform (STFT) [2].

[2]V_t^M is produced by projecting the current and past depth images $V_{1\ldots t}^D$ to the egocentric ground plane and stitching together these projections into a cumulative map of the scene.

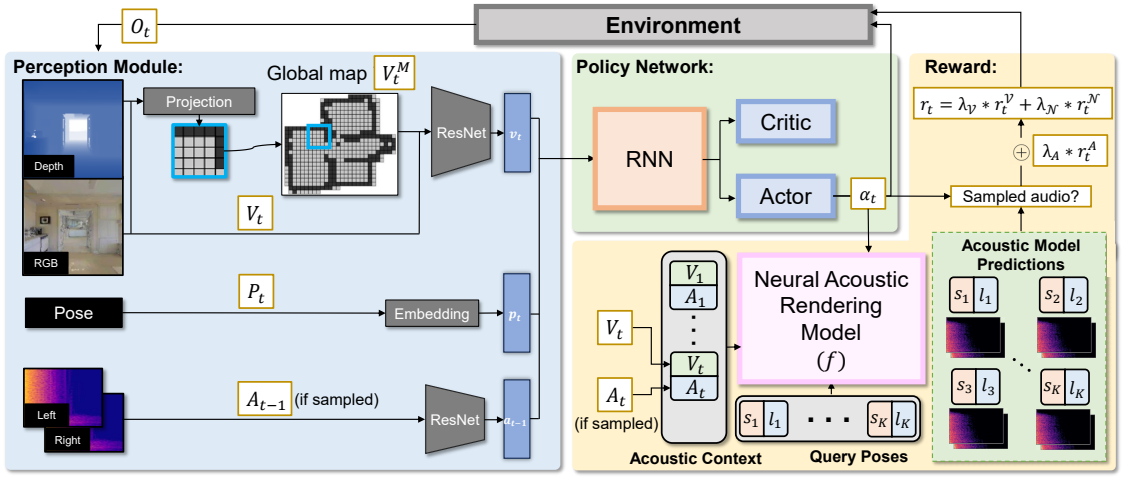

Fig. 2. **ActiveRIR policy network architecture and reward.** At each step t, our policy π receives an egocentric visual input V_t, the camera pose P_t, and the binaural echo response A_{t-1}—if it was sampled by the policy at the previous step—and predicts an action α_t that decides both how the agent should move, and if it should sample the current echo response A_t. It then uses A_t along with the current visual input V_t to improve its acoustics prediction quality. Given these audio-visual samples ("Context") collected over an episode, the agent uses an off-the-shelf acoustic rendering model to predict the RIR for any arbitrary query pair of sound source and receiver locations. We train our policy with an audio-visual reward which encourages healthy exploration of the scene in search of acoustically important locations, and guides the agent *when* to sample highly valuable observations, subject to a maximum audio sample budget.

features v_t and pose features p_t. The visual and audio encoders are ResNets [34], and the pose encoder is a sinusoidal positional embedding [35]. Next, we concatenate a_{t-1}, v_t and p_t into o_t and feed it to the policy network, which consists of an RNN and an actor-critic module. The RNN estimates an updated history h_t along with the current state representation g_t, using the fused feature o_t and the history of states h_{t-1}. The actor-critic module takes g_t and h_{t-1} as inputs and predicts a policy distribution $\pi_\phi(\alpha_t|g, t, h_{t-1})$ along with the value of the state $H_\phi(g_t, h_{t-1})$. Finally, the policy samples an action α_t from its action space $\mathcal{A}$ (*c.f.* Sec. III) per the distribution π_ϕ.

2) Policy reward: We propose a novel RL reward to train our policy π:

$$r_t = \lambda_A * r_t^A + \lambda_\mathcal{V} * r_t^\mathcal{V} + \lambda_\mathcal{N} * r_t^\mathcal{N}. \quad (1)$$

Here r_t^A is our novel *Acoustic Prediction* reward, which measures improvement in the environment acoustic model from the previous step $t-1$ to the current step t if audio was sampled, and is given by

$$r_t^A = \mathcal{L}_{t-1}^R - \mathcal{L}_t^R$$

, where $\mathcal{L}_t^R$ is the mean L1 distance between the predicted and ground-truth RIR magnitude spectrograms for a fixed set of K query positions selected from throughout the (training) scene, at step t. r_t^A is zero at the steps where the policy decides not to sample; thus it encourages the agent to sample an acoustic observation *when the agent is at a position that will improve the global acoustic prediction quality.*

Since this sparse reward can impact the stability of RL training, we augment r_t^A with an area-coverage reward [36] $r_t^\mathcal{V} = (\mathcal{V}_t - \mathcal{V}_{t-1})/\mathcal{V}_{t-1}$ that measures the relative increase in area coverage over time, where $\mathcal{V}_t$ and $\mathcal{V}_{t-1}$ are the total area covered by the agent at steps t and $t-1$, respectively. We also add a novelty reward [12], [16], [11] $r_t^\mathcal{N} = \frac{1}{\sqrt{n(m_t)}}$, where $n(m_t)$ is the visitation count at 1×1 meter discretized floorplan grid cell m_t. $\lambda_\mathcal{V}$ and $\lambda_\mathcal{N}$ are the respective reward weights. Whereas $r_t^\mathcal{V}$ rewards the agent for taking actions that expose new areas of the scene to the agent, $r_t^\mathcal{N}$ incentivizes increasing the visitation count of novel locations in the scene. $r_t^\mathcal{V}$ and $r_t^\mathcal{N}$ together promote a healthy exploration of the scene, which is crucial for learning a good policy, and also encourage stable RL training owing to their dense nature. We train π using Decentralized Distributed PPO (DD-PPO) [37]. The DD-PPO loss consists of a value loss, policy loss, and an entropy loss for further improving exploration.

B. Acoustic rendering model

Our sampling policy design is agnostic of the design of the acoustic rendering model f, providing the flexibility of using our policy with any off-the-shelf rendering model that can take audio-visual samples as inputs and predict the RIR R^Q at a source-receiver query Q in the scene (*c.f.* Sec. III). We use FewShot-RIR (FS-RIR) [2] as our rendering model backbone, due to its state-of-the-art performance. We also evaluate ActiveRIR's ability to generalize in experiments with a second backbone, NAF [3] (*c.f.* Sec. 5). Given audio-visual samples from a scene, FS-RIR uses a transformer [35] encoder to build an acoustic context of the scene, followd by a transformer-decoder to predict the RIR for an arbitrary source-receiver query using the acoustic context. We pre-train the acoustic rendering model f in a disembodied fashion—without a sampling policy—randomly selecting observations from a scene to add to the acoustic context. Next, we train our policy

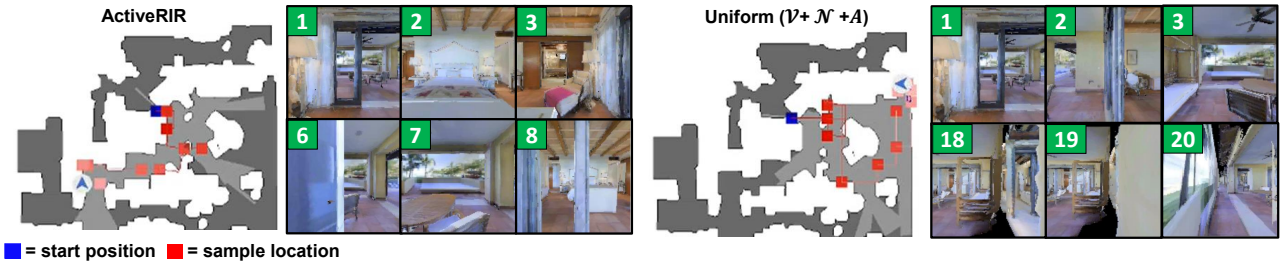

Fig. 3. **ActiveRIR vs. Uniform sampling.** The ActiveRIR agent (**far left**) navigates an environment and actively samples observations, collecting context **(left)** from regions of the environment where acoustics rapidly change—such as in a winding hallway—and which are visually *and* acoustically distinct from other samples in the context. In contrast, an agent passively sampling at a uniform interval **(right)** collects an acoustic context **(far right)** with spatial and visual redundancy, as observed by the bottom two images which show similar views of the same room captured only 1 meter apart.

with RL while keeping f frozen. This helps improve RL training stability by ensuring stationarity in the reward distribution (*c.f.* Sec. IV-A).

V. Experiments

A. Experimental setup

We use the state-of-the-art SoundSpaces (SS) 2.0 [14] acoustics simulator, built on top of the AI-Habitat [38] simulator and Matterport3D (MP3D) [15] scenes. MP3D consists of photorealistic scans of diverse, real-world multi-room indoor environments complete with furniture and other objects (*e.g.* tables/sofas/desks/lamps). SS 2.0 supports continuous rendering of precise spatial audio in arbitrary 3D scenes, capturing all major acoustic phenomena [14]. We use 78 diverse MP3D scenes, split into train/val/test sets of 56/10/12, preserving diverse scene types and sizes within each split. We train our policy for 150K PPO iterations and validate/test on 123/225 inference episodes respectively sampled in proportion to scene sizes within each split.

We place the agent at a random location in a scene at the start of every episode. We set $T = 200$ to account for the minimal time needed to traverse an average scene average in Matterport3D [15], and set the audio sample budget to $N = 20$ samples to match the context size used in FS-RIR [2]. We set $\lambda_A = 2{\times}10^5$, $\lambda_\mathcal{V} = 2{\times}10^2$ and $\lambda_\mathcal{N}$ =10.0 to place the component rewards on the same scale. The agent has turn and movement resolution 90° and 1 meter respectively. We evaluate performance with **STFT L1 Error (STFT)** [2], [19], which measures mean L1 error between predicted and ground-truth RIR magnitude spectrograms at $K = 60$ global query poses per scene, sampled randomly at the scene level.

B. Baselines

Random agent: an agent that chooses an action from our action space $\mathcal{A}$ randomly. **Forward agent**: an agent that only moves forward, sampling observations uniformly (every $\frac{T}{N}$ steps). **Greedy agent:** an agent that navigates randomly, and greedily selects the first N observations on its path.

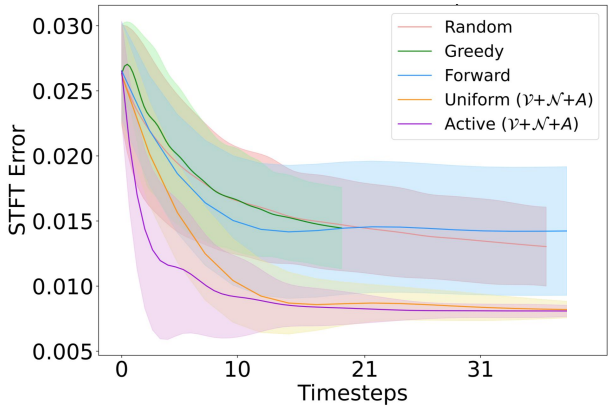

Fig. 4. **Acoustic prediction quality vs timesteps.** ActiveRIR (purple) rapidly minimizes STFT error in the acoustic model in fewer steps than an acoustic agent sampling at a fixed interval (orange), and outperforms heuristic approaches as well.

VI. Results

Table I shows acoustic prediction performance on unseen test environments. ActiveRIR significantly outperforms naive approaches (Greedy, Forward, Random). Fig. 4 displays STFT error as a function of the agent's steps. ActiveRIR (purple) efficiently navigates toward and captures valuable acoustic observations that rapidly minimize STFT error in the acoustic model, compared both to a passive-sampling audio-visual agent which samples every $\frac{T}{N}$ steps ("Uniform", orange) as well as an array of heuristic approaches.

1) Model analysis:

a) Reward variations: To evaluate the impact of our Acoustic Prediction reward, we train active policies with ablations of our audio-visual reward (Table II). We outperform Coverage ($\mathcal{V}$) and Novelty ($\mathcal{N}$) agents as well as the Exploration agent ($\mathcal{V}$+$\mathcal{N}$), which was trained with only the spatio-visual component of our audio-visual reward (*c.f.* Sec. IV-A.2), validating the importance of acoustic information in the agent's decision to *Sample* an observation beyond what can be inferred from visual and/or spatial sensory information alone. We also significantly outperform a passive audio-visual agent (Uniform), confirming that actively determining *when* to sample plays a critical role in collecting valu-

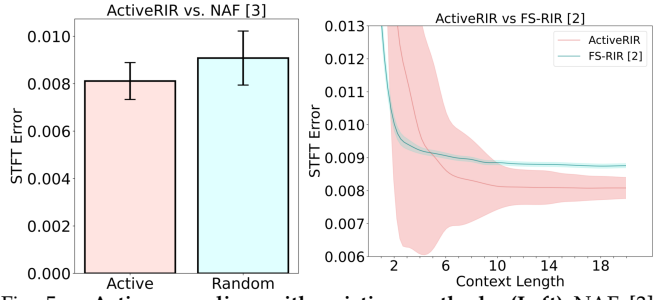

Fig. 5. **Active sampling with existing methods. (Left)** NAF [3] trained on ActiveRIR-collected context outperforms NAF trained on context collected by a random policy, demonstrating ActiveRIR's ability to select valuable acoustic context agnostic of the acoustic rendering model. **(Right)** As we grow the context size, ActiveRIR samples high-value observations that rapidly improve global scene acoustic error, producing a final acoustic model with significantly lower error than FS-RIR [2].

able acoustic context beyond simply navigating *towards* acoustically and visually novel areas. Inspired by [33], we evaluate a local variant of our acoustic reward, defined as the improvement in the RIR prediction error for a query closest to the agent's current position. We outperform this local acoustic agent, demonstrating that our global acoustic reward helps collect samples most valuable for *global* acoustic model.

b) Pose sensor noise: We evaluate ActiveRIR's robustness to Gaussian noise in pose and actuation. STFT error only increases from 8.08×10^{-3} to 8.10×10^{-3}, highlighting the effectiveness of our design choices.

2) Comparison with SOTA: Fig. 5 compares ActiveRIR to SOTA methods without embodiment and navigation budget. First, we compare against FS-RIR [2], which collects samples randomly throughout the environment (Fig. 5, right). Despite initial progress, FS-RIR plateaus as the episode progresses, while ActiveRIR continues selecting highly-informative samples that further reduce STFT error. Importantly, ActiveRIR produces a final acoustic model with significantly lower error than FS-RIR, *despite FS-RIR's prior knowledge of the floorplan for randomly sampling observations*. Furthermore, to achieve the error of ActiveRIR, FS-RIR needs to acquire a context length of 86 samples, more than 4x the samples required for our model. While randomly sampling from throughout the space may outperform our policy in high-resource regimes, ActiveRIR strongly outperforms FS-RIR in cost-efficient settings with restrictive time or sample budgets—which are practical considerations when considering physical constraints on access/time spent in the space, and the intrusive nature of emitting a sound impulse.

To determine the value of acoustic context collected by ActiveRIR agnostic of the acoustic rendering model f, we train two Neural Acoustic Field (NAF) [3] models on the test scenes using 1) ActiveRIR-collected context and 2) context collected by a random policy, and evaluate mean STFT error across 100 query poses sampled randomly from throughout the scene (Fig 5, left). The ActiveRIR-trained NAF consistently outperforms

Policy	STFT Error ↓
Greedy	14.46
Forward	14.68
Random	13.04
ActiveRIR ($\mathcal{V}$+$\mathcal{N}$+A)	**8.08**

TABLE I

Acoustic prediction performance. Reported in base 10^{-3}.

Reward	Sampling	STFT Error ↓
$\mathcal{V}$+$\mathcal{N}$+A	Uniform	8.24
$\mathcal{V}$	Active	8.15
$\mathcal{N}$	Active	8.11
$\mathcal{V}$+$\mathcal{N}$	Active	8.15
$\mathcal{V}$+A	Active	8.21
$\mathcal{N}$+A	Active	8.15
$\mathcal{V}$+$\mathcal{N}$+A (local)	Active	8.11
$\mathcal{V}$+$\mathcal{N}$+A	Active	**8.08**

TABLE II

Reward variations in ActiveRIR. Reported in base 10^{-3}.

NAF trained on the random policy's collected context, demonstrating that ActiveRIR collects a rich, generalizable acoustic context and is flexible with respect to the acoustic rendering model.

3) Qualitative analysis: Fig. 3 visualizes the importance of active sampling. ActiveRIR captures audio-visual samples in areas where acoustics can dynamically change, such as winding hallways (far left), while also ensuring that the context is spatially and visually distinct, as can be observed by the diverse views in the active context (left). In contrast, a passive agent (right) does not actively avoid capturing redundant observations (bottom two images), and collects samples in the main cavity of the scene where acoustics are relatively stationary between nearby poses. Also see submitted video.

VII. CONCLUSIONS

We propose a new task, *active acoustic sampling*, in which an agent must navigate an unmapped environment and collect audio-visual samples to construct a model of the scene's acoustics within a given time and sample budget. We propose an active sampling policy trained with a novel audio-visual exploration reward which guides an agent to navigate towards and select high-value audio-visual samples that yield a high-quality acoustic model. Our policy outperforms passive approaches in >70% fewer timesteps, and outperforms robust visual and spatial exploration agents as well as SOTA environment acoustic modeling methods [2] across diverse indoor scenes. We show that the performance gain using ActiveRIR-collected samples generalizes across multiple acoustic rendering models, demonstrating promising potential for ActiveRIR to improve existing acoustic rendering methods. In future work, we plan to explore 3D scene reconstruction from acoustic exploration.

References

[1] V. Välimäki, J. Parker, L. Savioja, J. Smith, and J. Abel, "More than 50 years of artificial reverberation," in *Proc. 60th International Conference of the Audio Engineering Society*, S. Goetze and A. Spriet, Eds. United States: Audio Engineering Society, 2016, aES International Conference on Dereverberation and Reverberation of Audio, Music, and Speech, DREAMS ; Conference date: 03-02-2016 Through 05-02-2016.

[2] S. Majumder, C. Chen, Z. Al-Halah, and K. Grauman, "Few-shot audio-visual learning of environment acoustics," 2022.

[3] A. Luo, Y. Du, M. J. Tarr, J. B. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," 2023.

[4] A. Ratnarajah, Z. Tang, R. C. Aralikatti, and D. Manocha, "Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes," *arXiv preprint arXiv:2205.09248*, 2022.

[5] A. Ratnarajah and D. Manocha, "Listen2scene: Interactive material-aware binaural sound propagation for reconstructed 3d scenes," 2024.

[6] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "Fast-rir: Fast neural diffuse room impulse response generator," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 571–575.

[7] C. Chen, S. Majumder, Z. Al-Halah, R. Gao, S. K. Ramakrishnan, and K. Grauman, "Learning to set waypoints for audio-visual navigation," 2021.

[8] S. Majumder and K. Grauman, "Active audio-visual separation of dynamic sound sources," 2022.

[9] S. Majumder, Z. Al-Halah, and K. Grauman, "Move2hear: Active audio-visual source separation," 2021.

[10] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu, "Neural acoustic context field: Rendering realistic room impulse response with neural fields," 2023.

[11] S. K. Ramakrishnan, D. Jayaraman, and K. Grauman, "An exploration of embodied visual exploration," 2020.

[12] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, "Unifying count-based exploration and intrinsic motivation," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.

[13] A. L. Strehl and M. L. Littman, "An analysis of model-based Interval Estimation for Markov Decision Processes," *Journal of Computer and System Sciences*, vol. 74, no. 8, pp. 1309–1331, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022000008000767

[14] C. Chen, C. Schissler, S. Garg, P. Kobernik, A. Clegg, P. Calamia, D. Batra, P. W. Robinson, and K. Grauman, "Soundspaces 2.0: A simulation platform for visual-acoustic learning," 2023.

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

[16] N. Savinov, A. Raichuk, R. Marinier, D. Vincent, M. Pollefeys, T. Lillicrap, and S. Gelly, "Episodic curiosity through reachability," in *International Conference on Learning Representations (ICLR)*, 2019.

[17] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori, "Image2reverb: Cross-modal reverb impulse response synthesis," 2021.

[18] L. Remaggi, H. Kim, P. J. B. Jackson, and A. Hilton, "Reproducing real world acoustics in virtual reality using spherical cameras," 2019.

[19] C. Chen, R. Gao, P. Calamia, and K. Grauman, "Visual acoustic matching," 2022.

[20] R. Gao and K. Grauman, "2.5d visual sound," 2019.

[21] A. Somayazulu, C. Chen, and K. Grauman, "Self-supervised visual acoustic matching," 2023.

[22] X. Xu, H. Zhou, Z. Liu, B. Dai, X. Wang, and D. Lin, "Visually informed binaural audio generation without binaural audios," 2021.

[23] K. K. Rachavarapu, A. Aakanksha, V. Sundaresha, and R. A. N, "Localize to binauralize: Audio spatialization from visual sound source localization," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1910–1919.

[24] N. Raghuvanshi and J. Snyder, "Parametric Wave Field Coding for Precomputed Sound Propagation," *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2014*, vol. 33, July 2014. [Online]. Available: https://www.microsoft.com/en-us/research/publication/parametric-wave-field-coding-precomputed-sound-propagation/

[25] R. Mehra, L. Antani, S. Kim, and D. Manocha, "Source and listener directivity for interactive wave-based sound propagation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 4, pp. 495–503, 2014.

[26] C. R. A. Chaitanya, N. Raghuvanshi, K. W. Godin, Z. Zhang, D. Nowrouzezahrai, and J. M. Snyder, "Directional sources and listeners in interactive sound propagation using reciprocal wave field coding," *ACM Trans. Graph.*, vol. 39, no. 4, aug 2020. [Online]. Available: https://doi.org/10.1145/3386569.3392459

[27] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," 2020.

[28] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," 2021.

[29] Y. Yu, W. Huang, F. Sun, C. Chen, Y. Wang, and X. Liu, "Sound adversarial audio-visual navigation," 2022.

[30] S. Purushwalkam, S. V. A. Gari, V. K. Ithapu, C. Schissler, P. Robinson, A. Gupta, and K. Grauman, "Audio-visual floorplan reconstruction," *arXiv preprint arXiv:2012.15470*, 2020.

[31] S. Majumder, H. Jiang, P. Moulon, E. Henderson, P. Calamia, K. Grauman, and V. K. Ithapu, "Chat2map: Efficient scene mapping from multi-ego conversations," 2023.

[32] X. Hu, S. Purushwalkam, D. Harwath, and K. Grauman, "Learning to map efficiently by active echolocation," 2023.

[33] Y. Yu, C. Chen, L. Cao, F. Yang, and F. Sun, "Measuring Acoustics with Collaborative Multiple Agents," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, E. Elkind, Ed. International Joint Conferences on Artificial Intelligence Organization, Aug. 2023, pp. 335–343. [Online]. Available: https://doi.org/10.24963/ijcai.2023/38

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[36] T. Chen, S. Gupta, and A. Gupta, "Learning exploration policies for navigation," 2019.

[37] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," *arXiv preprint arXiv:1911.00357*, 2019.

[38] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9339–9347.