SONICVERSE: A Multisensory Simulation Platform for Embodied Household Agents that See and Hear

Ruohan Gao*, Hao Li*, Gokul Dharan, Zhuzhu Wang, Chengshu Li, Fei Xia
Silvio Savarese, Li Fei-Fei, Jiajun Wu
Stanford University

***Abstract*— Developing embodied agents in simulation has been a key research topic in recent years. Exciting new tasks, algorithms, and benchmarks have been developed in various simulators. However, most of them assume deaf agents in silent environments, while we humans perceive the world with multiple senses. We introduce SONICVERSE, a multisensory simulation platform with integrated audio-visual simulation for training household agents that can both see and hear. SONICVERSE models realistic continuous audio rendering in 3D environments in real-time. Together with a new audio-visual VR interface that allows humans to interact with agents with audio, SONICVERSE enables a series of embodied AI tasks that need audio-visual perception. For semantic audio-visual navigation in particular, we also propose a new multi-task learning model that achieves state-of-the-art performance. In addition, we demonstrate SONICVERSE's realism via sim-to-real transfer, which has not been achieved by other simulators: an agent trained in SONICVERSE can successfully perform audio-visual navigation in real-world environments. Sonicverse is available at: `https://github.com/StanfordVL/Sonicverse`.**

I. INTRODUCTION

Future household robots should be able to see, hear, and follow voice instructions from humans. They should find their way around your home, go where you need them, and fetch the desired object at your command. For example, a robot should be able to locate the owner by its voice and follow the speaker, go to the kitchen to attend to an accident when hearing a cracking sound, and deliver a bottle of water when hearing your request from the bedroom.

An increasing amount of work in embodied AI has thus studied visual navigation, where an agent must use its egocentric visual stream to intelligently move around, explore a new space that has not been mapped before, and navigate to its goal. The goal can be specified by a point [1], [2], [3], [4], an object [5], [6], [7], or a room [8]. Many state-of-the-art simulators [2], [9], [10], [11], [12], [13], [14], [15], [16], [17] are also designed for developing embodied AI agents to tackle various challenging tasks.

Despite the encouraging progress, there is a salient missing ingredient—the environment is silent, and the agents cannot hear. Limited prior work has tackled embodied learning with audio [16], [18], [19], [20], [21], [22]: SoundSpaces [16] and ThreeDWorld [13] are two notable exceptions. However, SoundSpaces uses a separate audio dataset of pre-computed room impulse responses at a discrete grid of spatial locations with a pre-defined height, which prevents sampling data at new locations; ThreeDWorld supports continuous-space audio rendering, but it assumes a box-shaped approximation for modeling 3D environments, which limits its realism.

We introduce SONICVERSE, a new multisensory simulation platform for training audio-visual embodied agents that overcome these limitations. Not only can SONICVERSE render audio over a continuous space in real-time, but it also achieves high fidelity spatial audio rendering by using the complete scene geometry and material properties. As shown in Figure 1, we can attach semantically meaningful sounds to existing object assets or rooms in a 3D environment (e.g., water tap sound in the kitchen), and have agents serve as listeners who can receive the transmitted audio in space. Furthermore, we also support audio streaming with a Virtual Reality (VR) interface, enabling many potential applications for voice-driven human-robot interaction. For example, we can instantiate a person wearing the VR headset as an audio-visual avatar in the simulated environment. The person can issue a voice command, and the robot can locate its position from the spatial cues of the binaural audio it receives, and navigate to the person as instructed.

As a case study, we tackle the semantic audio-visual navigation task with continuous audio, and use our SONICVERSE simulation platform as a testbed for training navigation agents. We propose a multi-task learning framework for both audio-visual navigation and occupancy map prediction. The goal of occupancy map prediction is to infer occupancy (free or occupied) for unseen regions based on the audio-visual context. Our key insight is that these two tasks are mutually beneficial: the better the audio-visual state feature learned for navigation, the more useful it is for occupancy prediction; the prediction of occupancy in turn regularizes the learning for audio-visual navigation. We also train a model for audio-goal navigation and successfully deploy agents trained in simulation to real-world environments.

Our contributions are threefold. First, we introduce SONICVERSE, a new multisensory simulation platform that models continuous audio rendering in 3D environments in real-time, providing a new testbed for many embodied AI and human-robot interaction tasks that need audio-visual perception. Second, we introduce a multi-task learning framework for semantic audio-visual navigation and occupancy map prediction, which achieves state-of-the-art results. Third, we are the first to show that audio-visual navigation agents trained in simulation can be successfully deployed in real-world environments.

*Equal contribution, in alphabetical order

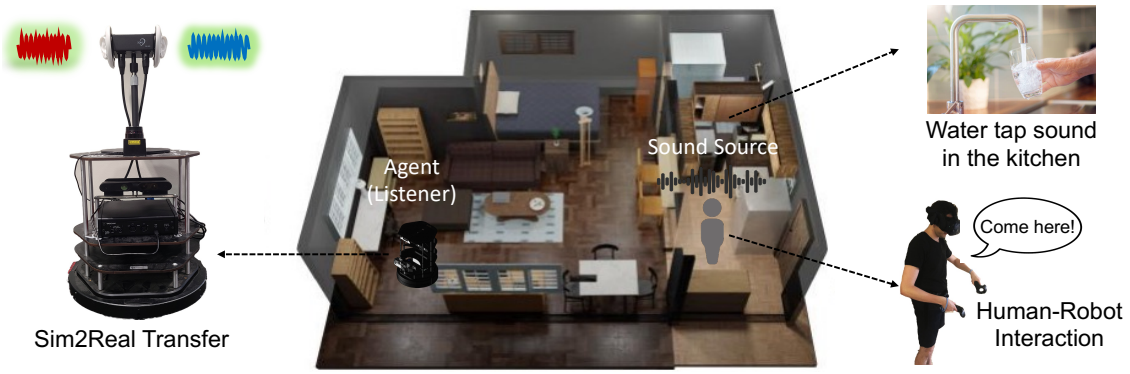

Fig. 1: Illustration of our SONICVERSE simulation platform. We visualize the top-down map of a 3D environment we support for spatial audio rendering. The source audio can either be a sound coming from a semantically meaningful object or room (e.g., water tap sound in the kitchen), or a voice from a person wearing the VR headset. The agent in the environment can act as a listener to receive directional information about the sound source and perform tasks that require audio-visual perception. An audio-visual embodied agent trained in SONICVERSE's environments can also generalize to their real-world counterparts (a TurtleBot with a binaural microphone shown on the left).

II. Related Work

Embodied AI Simulators. A series of well-designed simulation environments and benchmarks have been introduced for embodied AI research in the past several years: Habitat [2], [23], iGibson [10], [17], AI2Thor [24], RoboTHOR [11], Sapien [9], Robosuite [14], VirtualHome [12], RLBench [15], Meta-World [25], etc. These simulators above have enabled many new tasks and possibilities for training and developing embodied agents. However, none of them support multisensory perception.

The simulators that are most related to ours are SoundSpaces [16] and ThreeDWorld [13], each with distinct features and limitations. SoundSpaces can only render sounds at a discrete grid of spatial locations where the precomputed room impulse responses are available; ThreeDWorld assumes a user-specified box-shaped approximation of rooms for rendering spatial audio, which limits its realism. Our SONICVERSE simulator can both render continuous audio in 3D spaces in real-time, and achieve high realism by using the complete scene geometry and surface material properties. Concurrent with our work, SoundSpaces 2.0 [26] also supports on-the-fly continuous audio rendering. While their platform focuses more on supporting visual-acoustic learning and simulation results, we are the first to show that agents trained in a simulator can be successfully deployed in real-world environments for audio-visual navigation.

Visual Navigation. Significant progress has been made in the realm of visual navigation in recent years. Given the visual sensory data from onboard sensors, an embodied agent is tasked to reach a point-goal [1], [2], [3], [4], object-goal [5], [6], [7], [27], or room-goal [8], or to follow language instructions [28], [29], [30]. Some methods [3], [31] perform end-to-end training with implicit memory structure to predict actions directly from pixels, whereas others leverage semantic priors [6] and explicit memory structure like metric or topological maps [30] to facilitate navigation. Most existing visual navigation research is conducted without audio, an important gap that SONICVERSE aims to fill in.

Embodied Audio-Visual Learning. Recent work uses both vision and audio for a variety of embodied AI tasks, such as audio-visual navigation [16], [18], [19], [20], floor plan reconstruction [21], representation learning [22], curiosity-driven exploration [32], [33], or object-centric learning [34], [35]. Our work offers a new testbed to support these embodied AI tasks that require audio-visual perception, and we also perform a case study on the audio-visual navigation task to demonstrate the usefulness and realism of our simulator.

Audio-Visual Learning from Videos. Videos inherently contain both sights and sounds. Recent inspiring work leverage both modalities for a variety of interesting tasks, including self-supervised representation learning [36], [37], [38], [39], audio-visual source separation [40], [41], [42], [43], [44], sound source localization in images [37], [45], [46], [47], and spatial audio generation [48], [49], [50]. Unlike any of them, our work aims to enable embodied AI tasks that require audio-visual perception.

III. The SONICVERSE Simulation Platform

We introduce SONICVERSE, a new embodied AI simulation platform that supports audio-visual perception. SONICVERSE is built on top of iGibson 2.0 [17], which is an open-source interactive simulation environment for fast visual and physics simulation with a focus on household tasks. We augment it with the powerful open-source spatial audio SDK[1] of Resonance Audio [51] to enable audio-visual perception of the agents. Below, we introduce the main components of audio simulation, the 3D environments, and the key features of our SONICVERSE simulation platform.

A. Acoustic Simulation

We begin with the main components of audio simulation in SONICVERSE. Please see Supp. video for a demonstration.

[1] https://github.com/resonance-audio

(a) Speaker following (b) Voice-driven object retrieval (c) Auditory orientation training for blind people

Fig. 2: Illustrations of task prototypes enabled by our audio-visual virtual reality interface. See Supp. video for demos.

Direct Sound. Direct sound represents the sound that travels from the source to the listener without being obstructed or reflected by the environment. This component is attenuated as the distance between source and listener increases, as the energy of direct sound falls exponentially with distance. We also render near-field effects to get the most physically correct amplitude boost (including low-frequency boost) of a near-field source.

Dynamic Occlusion. Physical obstructions between the source and the listener attenuate the sound rendered to the listener by means of an occlusion node in Resonance Audio, which blocks high frequencies more than low frequencies to mirror real-world occlusion effects. At each simulation step, we calculate the number of objects that lie on a ray between the source and the listener, and attenuate the received audio accordingly.

Early Reflections and Late Reverberations. We initialize a user-specified number of reverb probes in the scene. These probes are uniformly spread throughout the scene, and we run a pre-simulation reverb-baking process by raycasting from each probe and measuring the characteristics of the early and late ray reflections. These properties vary depending on the size and shape of the immediate area as well as the material properties of the surfaces in the scene. Since these properties are pre-computed for each probe and are not updated dynamically, we perform the reverb-baking process on a version of the mesh that contains only the static structures of the scene. During simulation, SONICVERSE uses the reverb probe that is closest to the listener and renders reverberation effects by taking advantage of the pre-computed rt60s. Early reflections are rendered on-the-fly by also taking into account the listener's position relative to the probe, using a box-shaped approximation of the room. This different treatment of simulating early reflections and late reverberations allows both realistic geometry-based spatial audio rendering and real-time performance.

Head-Related Transfer Functions (HRTFs). Humans physically locate sound sources by taking advantage of time and level differences between the sound perceived by each ear. We utilize the HRTFs that incorporate these effects for rendering realistic binaural audio for the listeners.

B. 3D Environments

Though built on top of iGibson 2.0, SONICVERSE is flexible and supports two datasets of 3D scenes: Matterport3D [52] and iGibson [10]. We discuss their characteristics and how we configure them for audio-visual simulation below.

Matterport3D [52]: We use 85 large Matterport3D scenes of real-world homes and other indoor environments with $517m^2$ of floor space on average. Since Matterport3D scenes are static, we use the entire scene when performing reverb-baking. To do so, we map the semantic mesh categories to Resonance Audio material types (e.g., "wall" maps to "concrete block, painted", "curtain" maps to "curtain, heavy"), which determine the acoustic properties of the room surfaces such as scattering and absorption coefficients. This enables realistic reverberation and early-reflection modeling when raycasting from points in the scene.

iGibson [10]: We use 15 fully interactive scenes of real-world homes with furniture and articulated objects. As objects in iGibson scenes are movable, we only use the static skeleton of the scene to perform reverb-baking. This involves mapping the walls and ceiling to "concrete block, painted", assigning windows to "glass", and floors to "wood panel".

C. Key Features

Next, we introduce the two key features of SONICVERSE: the audio-visual Virtual Reality (VR) interface and the capability of Sim2Real transfer. Then we highlight our main differences against the two state-of-the-art audio-visual simulators.

Audio-Visual Virtual Reality Interface. We augment the VR interface from iGibson 2.0 [17] with streaming audio support, which is compatible with major commercially available VR headsets through OpenVR [53]. It enables the embodiment of a person wearing the VR headset as an audio-visual avatar and allows an agent to hear human-voiced commands in VR, opening many possibilities for human-robot interaction tasks with audio-visual perception. In Fig. 2, we illustrate three task prototypes enabled by our audio-visual VR interface: 1) *Speaker following*, where the agent needs to follow a person's voice and move together with the target person; 2) *Voice-driven object retrieval*, where the agent needs to receive the voice command from a person, sense its location from the binaural audio cues, fetch the right object as instructed, and finally deliver the object to the person who issues the command; 3) *Auditory orientation training for blind people*, where a blind person wears the VR headset and practices auditory orientation perception by navigating to the sound source in virtual environments. See Supp. video for demos of these applications.

Sim2Real Transfer. We use a TurtleBot as the embodied agent in our simulator, and we set up its real-world counterpart as shown in Fig. 3. We mount a 3Dio FS binaural microphone on the TurtleBot and use a Tascam

Fig. 3: Sim2Real set-up of a TurtleBot equipped with a binaural microphone in a real-world environment.

audio interface to process the received audio. The TurtleBot is equipped with an Asus XTION PRO RGBD camera, and an onboard Intel NUC. This allows us to verify that our audio simulation is realistic, such that the policies learned in SONICVERSE can be transferred to real-world environments.

Comparing to State-of-the-Art Audio-Visual Simulators. In comparison to SoundSpaces [16], a key advantage of our simulator is that it integrates audio and visual simulation by having sounds attached to dynamic objects in the scene. For example, given a loudspeaker playing music in the scene, an agent moving the speaker would hear the audio source move concurrently. Moreover, our interactive scenes allow for dynamic occlusion, making the sound intensity increase or decrease in response to the opening or closing of a door, for example. Our simulator also renders audio over a continuous space, whereas SoundSpaces uses a grid of discrete rendering points throughout the scene. ThreeDWorld [13] also uses a built-in version of Resonance Audio in UNITY, though with a user-specified box-shaped approximation. Our implementation achieves higher realism by using the complete scene geometry and automatically mapped materials for reverb-baking. We also demonstrate sufficient audio rendering fidelity for successful Sim2Real transfer. ThreeDWorld does not support spatial audio rendering for VR, though it does support basic sound rendering without advanced spatialization. We do not directly model object impact sounds as in ThreeDWorld, but SONICVERSE supports the integration of existing multisensory object assets with pre-computed audio simulation [34], [35].

IV. TRAINING AUDIO-VISUAL EMBODIED NAVIGATION AGENTS IN SONICVERSE

SONICVERSE supports many embodied AI tasks that require audio-visual perception. We tackle the challenging *semantic audio-visual navigation* [20] task as a case study to demonstrate the usefulness of our simulator. This is a more challenging version of audio-goal navigation [16], [18], in which an agent must locate a consistently-sounding source. In semantic audio-visual navigation, objects make sounds consistent with their real-world counterparts (e.g., doors make creaking sounds), and these sounds only last for a short period of time. The agent must therefore be able to localize the sound source well after it has stopped emitting sound, perhaps by leveraging the learned knowledge about which objects can emit certain sounds.

Task Definition. In this task, an agent is required to navigate to a specific semantically meaningful object in an unseen and unmapped environment by hearing the sound emitted by that object. The sound can be non-periodic, discontinuous, and of varied length. To reach the target object, the agent has to reason about the semantic category of the sounding object as well as the binaural spatial cues from audio perception. We use a TurtleBot as the agent for our experiments. We use 15 semantically meaningful sounds used in [20], including the sounds from the sink, cushion, tv, shower, etc. Each sound is one-to-one mapped to a specific target category. To be considered successful, the agent needs to locate the target position even after the sound stops, and navigate to the specific target object that was making the sound instead of any objects within the category.

Action and Observation Spaces. In contrast to the task's existing specification [20], which uses a discrete set of fixed-step translations and rotations, we use a continuous action space over robot wheel velocities. This makes the task setting more realistic and challenging, and more applicable to real-world robotics settings. The agent's observations include an RGB image, a depth map, the binaural audio received at its two ears, a bump sensor input, and its current pose with respect to the starting location.

Episode Specification and Success Criterion. Each episode is defined by the following: the scene, the start position and orientation of the agent, the target category, a target object within the category, and eight positions sampled within one meter of the target object position, which are considered as the nearby locations that define the object boundary. The agent is considered to meet the success criterion when it reaches any of the nine terminal positions: the eight positions near the target and the original target object position. The distance tolerance for reaching the terminals is 0.36m, which is the width of the real TurtleBot.

Audio-Visual Navigation Model. We propose a multi-task learning framework to jointly learn for semantic audio-visual navigation and occupancy map prediction, as shown in Fig. 4. At each time step t, the agent receives egocentric visual observation consisting of an RGB image and a depth map, and binaural audio at the agent's left and right ears represented as a two-channel audio spectrogram. We extract the visual and audio features from the visual encoder and the audio encoder, respectively.

For semantic audio-visual navigation, we adopt the base architecture from SAVi [20], adapted from the scene memory transformer network [54]. It mainly consists of two components: 1) *Goal Predictor*, which takes the audio feature and the agent's current pose as input to predict a goal descriptor that contains information about the sound source location and the object category of the sound; and 2) *Audio-Visual Transformer*, which uses a memory module to encode the agent's observations and uses a self-attention mechanism to reason about the 3D environment seen so far. The decoder of the transformer takes the output of the goal predictor and the encoded observations in its memory, and predicts the state feature s_t, which is then fed to an actor-critic network for predicting the next action a_t. We follow the two-stage training paradigm in [20] using decentralized distributed

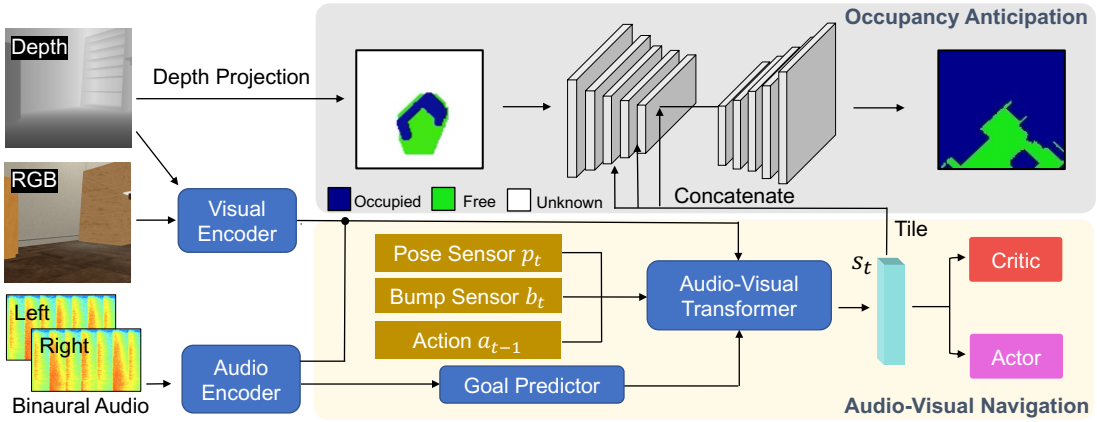

Fig. 4: Our multi-task learning framework for audio-visual navigation. We propose to jointly learn to predict the occupancy maps and the next action to reach the goal as the agent navigates in the environment. The two tasks are mutually beneficial. The better audio-visual state features s_t learned for navigation, the more useful they are for occupancy prediction; the prediction of occupancy in turn regularizes the training for audio-visual navigation.

proximal policy optimization [3].

For occupancy map prediction, we formulate it as a pixel-wise classification task, following [55]. We represent the egocentric occupancy as a top-down map $p \in [0,1]^{V \times V}$, which comprises a local area of $V \times V$ cells in front of the camera that represents a region of 5m $\times$ 5m. The value in each cell represents the probability of the cell being occupied. The ground-truth local occupancy is obtained by using the 3D meshes of the corresponding indoor environments. We use a U-Net [56] for occupancy map prediction. The input to the encoder is the local occupancy map obtained from depth projection by setting height thresholds on the point cloud obtained from depth and camera intrinsics [57]. We then replicate and tile the state feature vector to match the spatial dimension of the feature maps, and concatenate along the channel dimension for the last three layers of the encoder. The decoder then takes the fused feature map as input and outputs the predicted local occupancy map through a series of up-convolutional layers for both the visible and invisible cells. We use the binary cross-entropy loss for training the occupancy anticipation network.

Our occupancy map anticipation module is similar to prior methods in robotics and embodied visual navigation that build continuous representations of the world [58], [59], [60], [61], [55]. However, we jointly learn occupancy anticipation and audio-visual navigation, with the new insight that predicting accurate occupancy maps helps to learn better audio-visual features useful for navigation.

V. Experiments

We show our experiment results on audio-visual navigation, and how we transfer agents trained in our SoundSpaces simulator to real-world environments.

Baselines. We evaluate on audio-visual separation and compare to a series of baseline methods [16], [19], [18], [20]:

- Random Agent: A baseline that randomly samples an action at each time step and automatically stops after the agent reaches its goal.
- Gan et al. [18]: A map-based approach that predicts the goal location from audio and then has the agent navigate to the predicted location with an analytical path planner.
- Chen et al. [16]: An end-to-end approach for training RL navigation agents that leverage audio-visual observations, and it uses a GRU RNN to encode past memory.
- SAVi [20]: A state-of-the-art semantic audio-visual navigation model that uses a goal descriptor network to provide both location and object category information to the agent, and uses a transformer-based policy network.

Metrics. We evaluate using the following metrics [16], [20]: 1) success rate (SR), which is the fraction of successful episodes; 2) success weighted by (normalized inverse) Path Length (SPL) [62], which is success times the ratio of the agent's path length to the shortest path; 3) success weighted by the inverse of the number of actions (SNA), which penalizes collisions and in-place rotations. Results for all metrics are obtained by averaging over 1,000 test trails.

Semantic Audio-Visual Navigation. We use the iGibson dataset and the Matterport3D dataset for evaluation. On both datasets, we evaluate under two settings following the evaluation protocols from [16], [20]: 1) *heard* sounds—train and test on the same sound in unseen environments, and 2) *unheard* sounds—train and test on disjoint sounds in unseen environments. Table I shows the results. Our model significantly outperforms prior audio-visual navigation methods that do not take the semantic meaning of objects into consideration: Gan et al. [18] and Chen et al. [16]. This shows that our model successfully learns to match object categories with their sounds and leverage the semantic cues to navigate to the goal more efficiently. Compared to SAVi [20]—the state-of-the-art model for semantic audio-visual navigation, our multi-task learning framework consistently outperforms it across all metrics on both heard and unheard sounds, demonstrating the benefit of jointly learning to predict occupancy maps during navigation.

Fig. 5 shows the navigation trajectories on top-down

Model	*iGibson*						*Matterport3D*					
	Heard			*Unheard*			*Heard*			*Unheard*		
	SR	SPL	SNA	SR	SPL	SNA	SR	SPL	SNA	SR	SPL	SNA
Random Agent	2.4	2.4	1.2	2.4	2.4	1.2	1.5	1.5	1.2	1.5	1.5	1.2
Gan et al. [18]	12.1	9.2	8.9	5.0	3.9	2.9	5.9	3.8	3.5	4.1	3.5	2.7
Chen et al. [16]	41.2	39.0	9.7	24.2	22.0	4.8	18.1	17.1	5.2	10.0	9.4	2.3
SAVi [20]	53.0	47.4	9.9	43.5	37.9	9.3	27.9	26.8	7.32	22.4	20.6	4.7
Ours	**60.2**	**53.6**	**13.8**	**47.1**	**41.9**	**10.6**	**38.4**	**37.7**	**10.5**	**32.4**	**29.4**	**7.4**

TABLE I: Semantic audio-visual navigation results on the iGibson dataset and the Matterport3D dataset. Our proposed multi-task learning framework compares favorably against the closest competitor SAVi, the current state-of-the-art method. It outperforms all other baselines that do not consider the semantic meaning of the objects by a large margin.

Fig. 5: Navigation trajectories on top-down maps for semantic audio-visual navigation. Chen et al. [16]: The agent gets lost near the entrance of the dining room after the sound stops. SAVi [20]: The agent struggles to avoid the obstacles and takes a long path to reach the goal. Ours: The agent successfully avoids all obstacles and efficiently navigates to the target object.

maps in a challenging scene of our model and two baseline methods. The agent of Chen et al. [16] does not consider the semantic meaning of the sound source, and thus gets lost near the entrance of the dining room after the sound stops. The SAVi agent can successfully navigate to the goal, but it takes a much longer path compared to our method. Our agent can better sense the obstacles and the sound that travels through the door—potentially due to better audio-visual state features learned jointly with occupancy anticipation—and efficiently navigate to the target object (sink).

Sim2Real Transfer. To demonstrate the realism of our SONICVERSE simulator, we transfer the audio-visual navigation agents trained in simulation to real-world environments. To ease Sim2Real transfer, we train an AudioGoal navigation agent that only uses depth and audio for navigation, and an AudioPointGoal navigation agent that additionally takes as input a GPS pointer towards the target [16]. We use ROS's built-in GMapping SLAM algorithm to obtain the TurtleBot's pose as a replacement of the GPS used during training.

During our real-world deployment, we find three steps essential to reduce the Sim2Real gap for successful policy transfer. First, since the TurtleBot makes ego-noise during movement, we record the sounds of the robot moving in various environments, and mix the source sound with a random segment of the noise recording during training. Second, we also randomly vary the source gain as additional data augmentation due to the mismatch of the sound volume in simulation and real-world experiments. Third, we calibrate the depth camera to the range of 0.8m-3.5m to be consistent with the setting in simulation.

Table II shows the Sim2Real results in the bedroom shown in Fig. 3. We perform 10 test trials with different robot starting locations or sound source locations. We can see that both models transfer reasonably well to the real world, and AudioPointGoal has a higher success rate due to the additional target pointer that complements the audio signal. To our best knowledge, this is the first result that shows audio-visual navigation policies trained in simulation can be successfully transferred to real-world environments, demonstrating the realism of our multisensory simulation. See Supp. video for qualitative examples.

	SR	SPL	SNA
AudioGoal	30.0	22.2	5.5
AudioPointGoal	50.0	31.5	24.6

TABLE II: Sim2Real results for audio-visual navigation.

VI. CONCLUSION

We presented SONICVERSE, a multisensory simulation platform for training household agents that can both see and hear. Our simulator can render continuous audio in 3D environments in real-time and also support audio streaming with VR, providing a new testbed for many embodied AI tasks that need audio-visual perception. Using the audio-visual navigation task as a case study, we propose a new model for semantic audio-visual navigation, outperforming a series of prior methods. Furthermore, we successfully deploy agents trained in simulation into real-world environments. We look forward to the embodied multisensory learning research that will be enabled by SONICVERSE.

ACKNOWLEDGMENT

This work is in part supported by ONR MURI N00014-22-1-2740, NSF #2120095, Stanford Institute for Human-Centered AI (HAI), Adobe, Amazon, Bosch, Meta, and Salesforce.

REFERENCES

[1] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *CVPR*, 2017.

[2] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A Platform for Embodied AI Research," in *ICCV*, 2019.

[3] 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," in *ICLR*, 2020.

[4] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural slam," in *ICLR*, 2020.

[5] Y. Zhu, D. Gordon, E. Kolve, D. Fox, L. Fei-Fei, A. Gupta, R. Mottaghi, and A. Farhadi, "Visual Semantic Planning using Deep Successor Representations," in *ICCV*, 2017.

[6] D. S. Chaplot, D. Gandhi, A. Gupta, and R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *NeurIPS*, 2020.

[7] N. Savinov, A. Dosovitskiy, and V. Koltun, "Semi-parametric topological memory for navigation," in *ICLR*, 2018.

[8] X. Zhou, Y. Gao, and L. Guan, "Towards goal-directed navigation through combining learning based global and local planners," *Sensors*, 2019.

[9] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. J. Guibas, and H. Su, "SAPIEN: A simulated part-based interactive environment," in *CVPR*, 2020.

[10] B. Shen, F. Xia, C. Li, R. Martín-Martín, L. Fan, G. Wang, S. Buch, C. D'Arpino, S. Srivastava, L. P. Tchapmi, *et al.*, "igibson, a simulation environment for interactive tasks in large realisticscenes," in *IROS*, 2021.

[11] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford, L. Weihs, M. Yatskar, and A. Farhadi, "RoboTHOR: An Open Simulation-to-Real Embodied AI Platform," in *CVPR*, 2020.

[12] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba, "Virtualhome: Simulating household activities via programs," in *CVPR*, 2018.

[13] C. Gan, J. Schwartz, S. Alter, M. Schrimpf, J. Traer, J. De Freitas, J. Kubilius, A. Bhandwaldar, N. Haber, M. Sano, *et al.*, "Threedworld: A platform for interactive multi-modal physical simulation," in *NeurIPS Datasets and Benchmarks Track*, 2021.

[14] Y. Zhu, J. Wong, A. Mandlekar, and R. Martín-Martín, "robosuite: A modular simulation framework and benchmark for robot learning," *arXiv preprint arXiv:2009.12293*, 2020.

[15] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *RA-L*, 2020.

[16] 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," in *ECCV*, 2020.

[17] C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. Vainio, C. Gokmen, G. Dharan, T. Jain, *et al.*, "igibson 2.0: Object-centric simulation for robot learning of everyday household tasks," in *CoRL*, 2021.

[18] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *ICRA*, 2020.

[19] C. Chen, S. Majumder, A.-H. Ziad, R. Gao, S. Kumar Ramakrishnan, and K. Grauman, "Learning to set waypoints for audio-visual navigation," in *ICLR*, 2021.

[20] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in *CVPR*, 2021.

[21] S. Purushwalkam, S. V. A. Gari, V. K. Ithapu, C. Schissler, P. Robinson, A. Gupta, and K. Grauman, "Audio-visual floorplan reconstruction," in *ICCV*, 2021.

[22] R. Gao, C. Chen, Z. Al-Halab, C. Schissler, and K. Grauman, "Visualechoes: Spatial image representation learning through echolocation," in *ECCV*, 2020.

[23] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra, "Habitat 2.0: Training home assistants to rearrange their habitat," in *NeurIPS*, 2021.

[24] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

[25] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *CoRL*, 2020.

[26] 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," *arXiv*, 2022.

[27] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "Objectnav revisited: On evaluation of embodied agents navigating to objects," *arXiv preprint arXiv:2006.13171*, 2020.

[28] M. Zhao, P. Anderson, V. Jain, S. Wang, A. Ku, J. Baldridge, and E. Ie, "On the evaluation of vision-and-language navigation instructions," *arXiv preprint arXiv:2101.10504*, 2021.

[29] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *CVPR*, 2018.

[30] K. Chen, J. K. Chen, J. Chuang, M. Vázquez, and S. Savarese, "Topological planning with transformers for vision-and-language navigation," in *CVPR*, 2021.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] V. Dean, S. Tulsiani, and A. Gupta, "See, hear, explore: Curiosity via audio-visual association," in *NeurIPS*, 2020.

[33] C. Gan, X. Chen, P. Isola, A. Torralba, and J. B. Tenenbaum, "Noisy agents: Self-supervised exploration by predicting auditory events," *arXiv preprint arXiv:2007.13729*, 2020.

[34] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei, and J. Wu, "Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations," in *CoRL*, 2021.

[35] R. Gao, Z. Si, Y.-Y. Chang, S. Clarke, J. Bohg, L. Fei-Fei, W. Yuan, and J. Wu, "Objectfolder 2.0: A multisensory object dataset for sim2real transfer," in *CVPR*, 2022.

[36] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *ECCV*, 2016.

[37] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *ICCV*, 2017.

[38] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, 2018.

[39] B. Korbar, D. Tran, and L. Torresani, "Co-training of audio and video representations from self-supervised temporal synchronization," in *NeurIPS*, 2018.

[40] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *ECCV*, 2018.

[41] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *ECCV*, 2018.

[42] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba, "Music gesture for visual sound separation," in *CVPR*, 2020.

[43] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *ICCV*, 2019.

[44] E. Tzinis, S. Wisdom, A. Jansen, S. Hershey, T. Remez, D. P. Ellis, and J. R. Hershey, "Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds," in *ICLR*, 2021.

[45] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon, "Learning to localize sound source in visual scenes," in *CVPR*, 2018.

[46] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *ECCV*, 2018.

[47] D. Hu, R. Qian, M. Jiang, X. Tan, S. Wen, E. Ding, W. Lin, and D. Dou, "Discriminative sounding objects localization via self-supervised audiovisual matching," in *NeurIPS*, 2020.

[48] R. Gao and K. Grauman, "2.5d visual sound," in *CVPR*, 2019.

[49] P. Morgado, N. Vasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360° video," in *NeurIPS*, 2018.

[50] R. Garg, R. Gao, and K. Grauman, "Geometry-aware multi-task learning for binaural audio generation from video," in *BMVC*, 2021.

[51] M. Gorzel, A. Allen, I. Kelly, J. Kammerl, A. Gungormusler, H. Yeh, and F. Boland, "Efficient encoding and decoding of binaural sound with resonance audio," in *AES International Conference on Immersive and Interactive Audio*, 2019.

[52] 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," *3DV*, 2017.
[53] ValveSoftware, "Openvr," https://github.com/ValveSoftware/openvr.
[54] K. Fang, A. Toshev, L. Fei-Fei, and S. Savarese, "Scene memory transformer for embodied agents in long-horizon tasks," in *CVPR*, 2019.
[55] S. K. Ramakrishnan, Z. Al-Halah, and K. Grauman, "Occupancy anticipation for efficient exploration and navigation," in *ECCV*, 2020.
[56] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015.
[57] T. Chen, S. Gupta, and A. Gupta, "Learning exploration policies for navigation," *arXiv preprint arXiv:1903.01959*, 2019.
[58] S. T. O'Callaghan and F. T. Ramos, "Gaussian process occupancy maps," *The International Journal of Robotics Research*, 2012.
[59] F. Ramos and L. Ott, "Hilbert maps: Scalable continuous occupancy mapping with stochastic gradient descent," *The International Journal of Robotics Research*, 2016.
[60] K. Katyal, K. Popek, C. Paxton, P. Burlina, and G. D. Hager, "Uncertainty-aware occupancy map prediction using generative networks for robot navigation," in *ICRA*, 2019.
[61] R. Shrestha, F.-P. Tian, W. Feng, P. Tan, and R. Vaughan, "Learned map prediction for enhanced mobile robot exploration," in *ICRA*, 2019.
[62] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, *et al.*, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.