Finding Fallen Objects Via Asynchronous Audio-Visual Integration

Chuang Gan*, Yi Gu*, Siyuan Zhou, Jeremy Schwartz, Seth Alter, James Traer,
Dan Gutfreund, Joshua B. Tenenbaum, Josh H. McDermott*, Antonio Torralba*
MIT and MIT-IBM Watson AI Lab

Abstract

The way an object looks and sounds provide complementary reflections of its physical properties. In many settings cues from vision and audition arrive asynchronously but must be integrated, as when we hear an object dropped on the floor and then must find it. In this paper, we introduce a setting in which to study multi-modal object localization in 3D virtual environments. An object is dropped somewhere in a room. An embodied robot agent, equipped with camera and microphone, must determine what object has been dropped – and where – by combining audio and visual signals with knowledge of the underlying physics. To study this problem, we have generated a large-scale dataset – the Fallen Objects dataset – that includes 8000 instances of 30 physical object categories in 64 rooms. The dataset uses the ThreeDWorld Platform that can simulate physics-based impact sounds and complex physical interactions between objects in a photorealistic setting. As a first step toward addressing this challenge, we develop a set of embodied agent baselines, based on imitation learning, reinforcement learning, and modular planning, and perform an in-depth analysis of the challenge of this new task. This dataset is publicly available [1].

1. Introduction

Humans integrate multi-sensory data to understand the physical world around us. Consider the situation in which we hear the sound of an object falling somewhere in our house. What was it that fell, and where? Just by listening to the sound that is produced, we can usually determine not only the approximate location of the object that fell but also aspects of its physical make-up. The loudness of the sound, along with the reverberation that accompanies it, tells us much about the object's size, force of impact, and distance [53]. And we can usually tell whether the object was metal, wood or plastic [27,52], and whether it rolled on the floor after impact [2]. However, the sound alone is often not enough to precisely reveal the identity and location of the object, and instead must be used to guide visual search. Once we are in the vicinity of the fallen object, we use vision to find the object on the floor that is consistent with what we heard. We term this 'asynchronous audio-visual integration'.

Figure 1. Illustration of the proposed embodied physical sound source localization: an agent hears a physical object fall somewhere in the same room, and is required to find it via asynchronous audio-visual integration.

Replicating similar capabilities in assistive robots will be useful for many real-world applications. For example, robots may need to fetch a screw dropped in the middle of an automated production line which might shut down operations until the item is located and safely removed. The field of audio-visual navigation has made exciting progress by using both visual and audio modalities to drive embodied agents to navigate towards the target location of sound sources. Some of this earlier work [12, 20] defines the acoustic target as a repeating sound, such as a phone ringing or an alarm, providing a constant acoustic cue to the agent. However, in real-world situations, many sounds are intermittent, or are the result of a single non-repeating impact

*Equal Contribution
[1]Project page: http://fallen-object.csail.mit.edu

event. Recent work from Chen [11] made the first attempt to introduce semantic audio-visual navigation, in which objects in the environment must be localized via sound clips of short duration. The use of brief, non-repeating sounds was an advance over previous work, but the resulting dataset had two limitations. First, the platform that was used does not include impact sound synthesis capability, and therefore was unable to render the physical properties and interactions of objects via audio. The audio that was used did not reflect any underlying physical events such as objects falling, bouncing and possibly colliding with other objects, nor did it reflect the physical parameters of objects. As a result, the task could plausibly be solved primarily by learning semantic correlations between sounds and the scene context. Second, the locations of the sounding sources were defined in 2D uniform grids with fixed height, rather than the 3D locations that characterize sound source localization in real-world environments (in which the sounding objects could be anywhere in the room).

To remedy these limitations, we propose a new embodied AI challenge for multi-modal physical scene understanding. An embodied agent hears an unknown object fall to the ground, somewhere in the room it is in. The agent must then navigate within the room to locate and identify this physical object using both visual and auditory modalities (Figure 1). To support this task, we use the ThreeDWorld (TDW) simulation platform [19], which provides real-time synthesis of impact sounds, photo-realistic rendering and believable physical simulation. Audio is rendered with TDW's PyImpact Python library, which uses modal synthesis to generate audio from information about the material types and impact parameters of colliding objects (velocities, normal vectors and masses) [52]. Impact sounds are then spatialized using Resonance Audio, providing reverberation simulation that reflects the spatial dimensions of the room and the wall and floor materials being used.

We incorporated physical simulation of objects' physical interaction behavior and resulting impact sounds into embodied audio-visual navigation so as to pose several unique challenges for the agent. First, the diversity of locations of the target objects is increased relative to previous work, since the objects may be under or behind a sofa, on top of a cabinet, on a shelf, inside another containing object, or under a table. In particular, the agent is required to adjust its height in order to find the fallen object successfully. Second, we included distractor objects on floors, tabletops, and counter surfaces that disguise the location of target objects. Thus, the agent cannot simply leverage the correlation of the sound and scene to navigate towards the object. Instead, it must use the audio signal to infer the physical properties of the fallen object, constraining what this fallen object looks like and roughly where it has fallen.

We evaluate several agents on this benchmark. Though recent progress in embodied navigation has resulted in agents capable of navigating within an unfamiliar environment, experimental results suggest that it remains very challenging for these embodied agents to complete our task successfully. We believe models that perform well on our challenge will take a meaningful step towards more intelligent robots that could infer physical information about the scene from multi-sensory data. Our contributions are summarized as follows:

- We introduce a new *embodied physical sound source localization* task that aims to measure AI agents' physical inference abilities by finding fallen objects via asynchronous audio-visual integration.
- To support this challenge, we augment the existing TDW simulation platform by adding 64 new physical rooms and 8 new audio materials to simulate a diverse range of impact sounds in the different scenes.
- We create the *Fallen Objects* dataset, a comprehensive audio-visual dataset that includes over 8000 instances of 30 physical object categories fallen in 64 rooms. We will make this dataset publicly available.
- We develop several baseline agents as first steps to tackle this task. We use these baselines to perform in-depth analysis of the challenges presented by our benchmark, and also highlight potential directions to improve performances on the task.

2. Related Work

Embodied AI. A long-standing goal for the computer vision and robotics communities is to develop intelligent agents that could assist with everyday tasks. Since directly training robot agents in the real world is expensive, and makes it hard to benchmark different algorithms, there has been growing interest in using high-fidelity interactive 3D simulation to train and evaluate embodied agents. Representative platforms include Habitat [45], Gibson [57], i-Gibson [34], AI2-Thor [33], Virtual Home [40], Sapien [58], and ThreeDWorld [19]. These platform have also been used to generate large-scale datasets, making progress on pointing goal navigation [45,56], semantic navigation [10], interactive navigation [34], instruction following [48], audio-visual navigation [12, 20], rearrangement [7, 22, 49, 55], among other tasks. State-of-the-art approaches to embodied tasks include end-to-end training of neural policies using RL [56, 61] or hierarchical RL [6, 32] and modular approaches by integrating perception and mapping for path planning [9, 10, 28]. Recent work has also shown that web videos can be used to train embodied agents that could flexibly navigate to a goal [8, 29]. However, at present we lack datasets with which build multi-modal physical reasoning abilities into embodied agents. Our challenge aims to fill this gap.

Audio-Visual Navigation. Audio-visual navigation has recently emerged as a research focus in computer vision and robotics [11–13, 16, 20]. Unlike visual navigation, an audio sound source is used to define the target, and the agent must use both audio and visual signals to navigate to the target. Existing platforms used for this task simulate audio using a prerecorded repeating sound (*e.g.* alarm) [12, 20] or a brief semantically identifiable sound [11] downloaded from the Internet. Then they use acoustic reverberation simulations to calculate the impulse response at each of several pairs of sound emitter and receiver locations. Finally, they convolve the impulse response with the raw waveform to simulate the sound observation for the agent. As discussed above, this simplification is a useful first step towards incorporating audio signals into embodied AI, but it does not consider how the physical properties of objects, and the physics of their interactions, influence what the agent hears. By contrast, human listeners are believed to implicitly infer physical generative parameters from what they hear, and to use these inferences in estimating the distance of sound sources [53]. In this work, we remedy these limitations by incorporating new simulation techniques of physically-driven impact sound synthesis for complex object interactions in near photo-realistic environments. These enable us to create a large-scale audio-visual dataset of fallen objects that poses unique challenges for object-centric physical inference from multi-sensory data.

Audio-Visual Learning. Our work is also related to a body of research leveraging the relationship between vision and sound for model learning. A main theme of this previous work has been to show that the synchronization of audio and video can be utilized as a 'free' training signal. The applications include feature representation learning [4, 37, 39], visually-guided sound separation and localization [17, 25, 60], object localization [1, 5, 21], scene parsing [44, 51], active speaker detection [43], depth prediction [24], learning of scene structure [15], floor plan reconstruction [41], cross-modal generation [18, 26, 36, 38, 47, 50] and so on. Our work differs from these prior efforts in studying how to use asynchronous visual and audio information to perform physics-based tasks.

3. Fallen Objects Dataset

We built the *Fallen Objects* dataset to assess an agent's asynchronous audio-visual integration capabilities for physical object localization in simulated 3D environments that mimic the sensory and interactive richness of the real world. We chose to use the ThreeDWorld (TDW) platform [19] to generate the dataset because TDW supports both near-photo-realistic image rendering, physically-based sound rendering [52] with reverberation, and realistic physical interactions between objects and agents. In this section, we introduce the problem formulation and data collection process.

3.1. Problem Formulation

Task definition. An embodied agent with an egocentric-view camera and microphone hears an unknown object fall somewhere in the same room it is in (a study or kitchen room). The agent is then required to find which object has fallen and where, as efficiently as possible, using asynchronous audio-visual integration.

Observation and action space The observation space for this agent contains a first-person-view RGB image, depth map, egocentric binaural audio as well as the agent's current pose. The audio observation is only available at the very beginning, when the object falls and settles (*i.e.* step 0). The agent's action space comprises *move forward, rotate left, rotate right, look up, look down*, and *found*.

Success criterion. The agent succeeds on the task if and only if it positions itself within 2 meters of the target object [8], with the target object appearing within the first-person view when the agent executes the *found* command. This is a strict success criterion inspired by consideration of real-world applications, in which an assistive robot needs to explicitly understand the goal and whether it has reached it.

3.2. Dataset Generation

Simulating this type of large-scale physical object event data involves several challenges. First, it requires a simulation platform that can deeply integrate physics behavior, impact-based sound synthesis, and image rendering, while also supporting an embodied agent that can interact with the virtual world. Second, it requires a rich set of object assets and material types that can generate a comprehensive and diverse set of physical object falling events. We chose the ThreeDworld platform because its advanced multi-modal data synthesis, physical simulation capabilities, and physical object library provide solutions to these challenges.

Simulating fallen physical objects in 3D environments. To support the dataset generation, we created 8 new 3D models of kitchen and study room environments, to be used in the TDW simulation engine. For each room environment we varied the wall and floor materials and included one of 4 different furniture layouts, for a total of 64 room variants. All objects in these room environment scenes respond to physics. We also define a set of 30 object categories that will be dropped under the control of physics; these objects are assigned varying physical material properties such as mass, friction and restitution (bounciness). Objects are additionally assigned an audio material that matches their variation in visual material (e.g. wood, metal, ceramic etc.) For this challenge, we created 8 new audio material types,

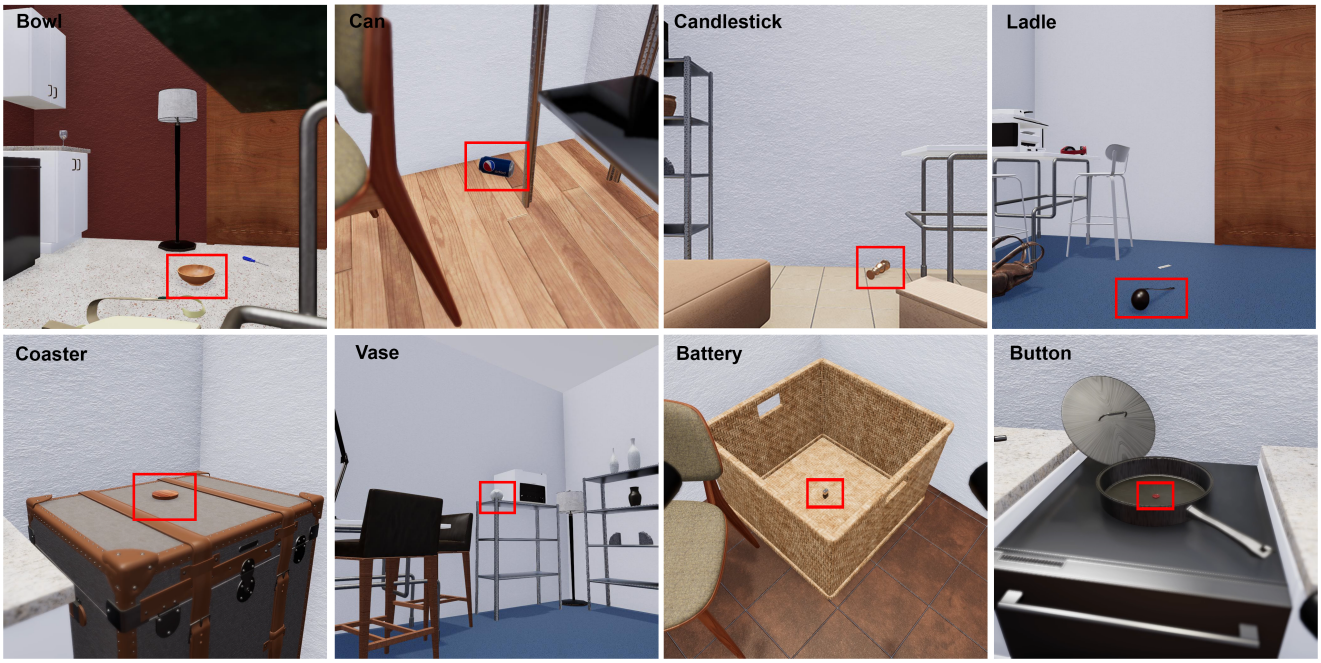

Figure 2. An overview of the fallen objects dataset. The objects and fall locations are diverse; objects like those depicted in the images can land on the floor, on a table or trunk, on top of a shelf, or hidden within containers such as a basket or a pan on the stove.

including both hard and soft plastics, rubber, stone and several additional wood materials, for a total of 14 materials. When objects collide with other objects or the floor, TDW's PyImpact module synthesizes impact sound data reflecting the object's physical and material properties as well as parameters of the collision such as velocity and angle of impact. Impact sounds are spatialized, providing real-time audio propagation, directional cues via head-related transfer function, and high-quality simulated reverberation that varies with room geometry and surface materials. Target object categories include: keys, coasters, cutlery, glasses, batteries, coins and corks, each using one of the 14 materials. The full list can be found in the supplementary material.

Procedural generation of impact sounds. For each room variant, we define 20 *fall zones*, i.e. locations where we expect the target objects to land when fallen. Fall zones can be on the floor, on top of tables, shelves or cabinets, or within open container objects such as boxes, baskets or pans on the stove. Fall zones can also contain additional "distractor" objects of similar size to the target object, as well as occluder objects that can hide it, both of which increase the difficulty of finding the target object. Some examples of fallen objects can be found in Figure 2.

To efficiently generate valid scenarios for this dataset, we developed a two-phase procedural generation pipeline: an initial rehearsal phase that generates guaranteed-valid fall event data, and a subsequent generation phase that uses that data to generate audio heard by the agent. Thus, before rendering any audio data, we removed any failed trial scenarios where objects fell outside of the desired fall zone.

1. **Rehearsal Phase.** We simulate target objects dropping onto their designated fall zone. Objects that fall or roll outside of their fall zone, and end up in plain view, are considered too easy for the agent to find and are filtered out. For each room variant, we initialize the scene and fall zone data. We then run a series of trials that: 1) define a randomized starting position and orientation above the fall zone for each target object; 2) drop the object under control of physics. All objects that remain within the drop zone when they cease moving are kept for the dataset. We record the start and end parameters for these objects (termed the 'fall event data') for the generation phase, described next.

2. **Generation Phase.** In this phase, we generate the actual audio produced by the fallen objects. For each room variant, we load the corresponding fall event data from the rehearsal phase, and initialize the reverberation and wall/floor material parameters. We then generate a series of trials, where each trial 1) spawns an embodied agent in a random free position and facing in a random direction; 2) adds the target object from the fall event and lets it fall; 3) generates a spatialized audio recording of the falling event, as heard by the agent, using the TDW sound engine.

Collecting demonstrations data of expert trajectories.

We further collect demonstration data for all scenarios to illustrate the shortest path and optimal action sequences to succeed in the task. We first calculate a shortest path to navigate from the agent's position to the fallen object based on the ground truth occupancy maps from the the simulation. After the agent arrives near the fallen object, we then adjust its orientation by executing actions *rotate left* or *rotate right* and its height by executing actions *look up* or *look down* to search for the target object. These expert trajectory demonstration data are intended to help researchers improve policy learning or incorporate heuristics into planning algorithms.

4. Experiments

4.1. Baselines

We implemented 5 baseline agents that include models using imitation learning, reinforcement learning and modular planning-based algorithms:

- **Decision Transformer [14]:** This is a recent state-of-the-art imitation learning framework, which leverages the transformer architecture [54] to model trajectories and predict future actions. There are 3 kinds of embeddings: state, action and return-to-go. We extract the features from RGBD images, sound spectrograms and segmentation masks as state embeddings and learn a linear layer for action and return-to-go embeddings. In addition, we learn timestep embeddings and add them to each token. We finally feed the last 20 timesteps tokens into the model and use a GPT model [42] to predict future actions. The model is trained using the expert trajectories provided as part of our dataset.
- **PPO [46]:** Similar to previous work [12], we train an end-to-end RL policy using Proximal Policy Optimization (PPO) by maximizing the reward of finding the audio goal (fallen objects in our case). This model takes the input history of RGBD images, segmentation masks as well as audio observations and outputs actions that the agent should execute in the environment. We adopt an oracle *found* action for this agent. Specifically, the agent will be considered successful if the fallen object is less than 2 meters from the agent while also appearing in the first-person-view image.
- **SAVi [11]:** This is a recent state-of-the-art framework for semantic audio-visual navigation. It first uses an auditory perception module to estimate an audio goal and then adopts a transformer-based deep RL framework that can learn to attend to the previous visual and audio observations to determine a path to the predicted audio goal. We use their open-source code to perform our task.
- **Modular Planning [10, 20]:** This is a hybrid and modular framework that integrates the auditory perception module used in [20] and the visual semantic mapping module used in [10]. The model first predicts the fallen object's category and location based on the audio, and then builds maps of the environment from RGBD images in search of the fallen object. If the agent does not find the goal object successfully during this process, it will adopt a frontier exploration strategy [59] to search unexplored areas. Otherwise, it will move close to the potential fallen object and find it.
- **Object-Goal [10]:** This is a recent state-of-the-art method for visual semantic navigation. It adopts a modular SLAM-based navigation approach. They first use depth projection and segmentation for semantic mapping and select a waypoint to explore. If the target object appears in the map, the agent plans a shortest path and navigates towards it. We use this baseline to test whether or not the agent needs audio signals to succeed on our task. To decide when to execute the *found* action, we provide 5 potential object categories to the agent (one of which is correct), which uses a heuristic approach to guess if an object is fallen or not. We used the open-source code from the original paper [10] and adapted it to our task.

4.2. Experimental Setup

Setup. We use 8000 instances of 30 physical object categories in 64 physically distinct rooms (32 study rooms and 32 kitchens) for the experiments. We used 6000 of these instances for training, 1000 for validation, and 1000 for testing. We report the models' performance on the test set. The agent received an audio observation only at the start of the trial. The agent then received an RGB-D observation at each subsequent time step. For the actions, we set *move forward* to 0.25m and *rotate* to 30 degrees. To test the generalization capabilities of these agents, we additionally carried out cross-scene generalization experiments described in section 4.5.

Evaluation metrics. We use three metrics to evaluate agents: Success Rate, Success weighted by Path Length (SPL) [3], and Success weighted by Number of Actions (SNA) [13].

The *Success Rate* is defined as the ratio of the number of times the agent successfully navigates to the target to the total number of test trials. A trial is considered successful based on four criteria: first, the agent needs to explicitly execute action *found*; second, the distance between the agent and the goal is less than 2 meter; third, the target physical object should appear in the agent's view; fourth, the length of the action sequence is less than the maximum number of allowed time steps N. We set N as 200 for all comparisons, as this was twice the average action sequence length of the demonstration data. The *SPL* is a metric that jointly considers the success rate weighted by the path length to reach the goal from the starting point. And the *SNA* jointly considers the number of actions and the success rate, by penalizing

collisions, rotations, and height adjustments needed to find the objects.

4.3. Implementation Details

For all the baselines, the input observation data are 300 $\times$ 300 size RGBD images, around 2 second binaural audio sampled at 44.1 kHz , and agent pose. Below we provide the details of the baseline models: the visual and auditory perception modules, semantic goal and occupancy map estimation, policy training, and the planners.

Visual and Auditory Perception Modules We first train a Mask-RCNN model [30] as a visual perception module for semantic segmentation of 30 pre-defined object categories. We then train an auditory perception module used for anticipating the category and location of the sounding fallen object in our environments. To pre-process the audio data, we transform the raw waveform into a spectrogram, which provides the input to a ResNet-18 network [31]. The network is trained to predict the object category from sound using a cross-entropy loss. To estimate the location of the fallen object, we follow previous work [20] by predicting the relative orientation and distances with respect to the agent, trained using a mean squared error (MSE) loss. The absolute coordinate of the estimated fallen object can then be calculated from the agent's coordinates and the relative location.

Semantic Goal and Occupancy Map. We represent the global top-down semantic goal map as $S_g \in \{0,1\}^{N\times N}$ and the occupancy map as $O_g \in \{0,1\}^{2\times N\times N}$, where N$\times$N represents the map size and each cell corresponds to a region of size 0.1m$\times$0.1m in the room. Each cell in the semantic goal map represents whether the corresponding region of the room contains a possible target object or not. For a region that might contain a possible target object, we map the geometric center of this semantic segmentation mask to a cell using the depth map. The two channels in the occupancy map represent whether the cell is occupied and explored, respectively. This is measured by whether the corresponding region contains an obstacle or has been observed. At each time step, the agent can recover a local semantic map $S_l \in \{0,1\}^S$ and a local occupancy map $O_l \in \{0,1\}^{2\times S}$ from the egocentric semantic segmentation and depth map of the visible area in front of the agent.

Planning. For planning, the agents first rely on the auditory perception module to infer a goal location and potential fallen object category from the sound. As the agents receive RGBD images at each time step, we run a visual perception module on the RGB images, returning segmentation masks of objects in the egocentric view. We then integrate the semantic segmentation with the depth image to construct an occupancy map and a semantic goal map for path planning. The planning module needs to correlate the semantic goal map with the object category inferred from the auditory perception module. The planner will provide a sequence of actions to move towards this target object along the shortest path if the target object is found around the position estimated from the audio. If the sound position prediction does not correspond to anything in the semantic goal map, the agent will use frontier exploration to navigate the unexplored area on the occupancy map in search of the fallen object.

Policy Training. For training PPO and SAVi agents, we use the same reward function for policy training. At each step, the agent receives a reward of $+1$ if it is close to the target location and -1 if it is far from it. The agent receives a reward of $+10$ if it finds the fallen object. There is also a -0.01 penalty for each time step.

For PPO, we first use CNNs to encode the RGBD images, the semantic segmentation and the spectrogram at each time step as feature vector and concatenate them as an input for the Gated Recurrent Unit (GRU) model. We train PPO with the Adam optimizer, using a learning rate of 2.5×10^{-4}. We collect the experience of 16 steps to update the policy. For SAVi, we use the same auditory perception module to initialize the weight of the target descriptor. The location prediction network is then fine-tuned with the ground truth locations online during policy training. We train SAVi with Adam using a learning rate of 2.5×10^{-4}. Follow the implementation in [11], we also roll out the policy for 150 steps and then use these experience data to update the policy every two epochs. We train both the PPO and SAVi models until convergence.

4.4. Result Analysis

Overall results. In Table 1, we report the overall performance of different agents on 1000 test scenes under the three evaluation metrics. The modular planning-based approach achieves the best results across all metrics but still fails in over half of the scenes. The RL model (PPO) performs poorly on the task, even when adopting an oracle *found*. Surprisingly, we observe that the previous best semantic audio-visual model [11] (SAVi) also struggles on this task. The reason might be that this model tends to capture the correlations between the scene context and sound. Our dataset tends to eliminate these kinds of shortcuts, because the physical object could fall anywhere in the rooms. The introduction of distractor objects might also pose an addition challenge for the cross-modality attention mechanism in [11]. The Decision Transformer achieves a success rate of 17%, which is comparable to that of the PPO. This result suggests that combining supervised learning from expert data with reinforcement learning or modular planning could be a promising direction for the future. We also notice that the modular SLAM-based framework for object-goal navigation could achieve around 22% success rate, much lower than that obtained using the audio signal to define the

Table 1. Navigation performance comparisons on the fallen object dataset. The modular planning-based agent achieves the best results. We also provide additional diagnosis results for this agent by replacing each individual component as an oracle module.

Method	Success Rate	SPL	SNA
Decision TransFormer [14]	0.17	0.12	0.14
PPO (Oracle found) [46]	0.19	0.15	0.14
SAVi [11]	0.23	0.16	0.10
Object-Goal [10]	0.22	0.18	0.17
Modular Planning [20]	**0.41**	**0.27**	**0.25**
Modular Planning (GT seg.)	0.47	0.34	0.32
Modular Planning (GT object)	0.67	0.41	0.38
Modular Planning (GT seg.+ object)	0.85	0.61	0.56
Modular Planning (GT seg.+ object + location)	0.93	0.83	0.80

Figure 3. Visualizations of the modular planning based agent. In each case, the top down view shows the agent's trajectory, while the numbered images show the agent's egocentric view: 1) at the start, 2) after identifying an object that looks like the fallen object, and 3) after navigating to the object and executing the *found* action. Row 1: fallen object is a pen; Row 2: fallen object is headphones.

goal location (41%). This indicates the essential need to incorporate multi-modal physical reasoning capabilities into this kind of modular planning-based algorithm to succeed on this task.

What Makes This Task Difficult? We also perform an in-depth analysis to understand the challenge of this dataset, using the best modular planning-based approach (Table 1). We leveraged the modular design of this method, replacing individual components of the modular framework with an oracle model and measuring the effect on performance. We performed this 'oracle swap' with the modules for visual perception (*i.e.* object segmentation), auditory object perception (*i.e.* object category predictions from sound), sound source localization (*i.e.* fallen object location), and the exploration strategy. We first implement two oracle agents with perfect visual perception (GT seg.) and auditory object perception (GT obj.), respectively. These experiments yielded two key observations. First, an oracle visual perception module improved the success rate marginally but significantly reduced the number of steps needed to find the target object. Second, an oracle auditory object category prediction model improved the success rate over 20%, indicating that one major bottleneck is accurately predicting object categories from audio alone. This could indicate that

Table 2. Evaluation of cross-scene generalization of different agents.

Method	Kitchen to Study Room			Study Room to Kitchen		
	Success Rate	SPL	SNA	Success Rate	SPL	SNA
PPO (Oracle found)	0.11	0.10	0.10	0.05	0.04	0.05
SAVi	0.20	0.11	0.09	0.19	0.14	0.11
Decision TransFormer	0.07	0.06	0.06	0.08	0.06	0.07
Object Navigation	0.18	0.14	0.13	0.15	0.14	0.13
Modular Planning	0.34	0.23	0.20	0.35	0.22	0.19

the trained auditory perception module failed to learn to optimally infer physics from sound. The inability to identify the fallen object category is also a major failure mode for all other baselines. One possible future approach to improve auditory recognition is to learn an object-centric neural implicit function [23] as a generative model to approximate an object's sound based on its physical properties, and to adopt an analysis-by-synthesis approach to invert the object physics from sound. When the two oracle visual and audio modules were combined (GT seg.+object), the agent performed well (around 86% success rate).

We also implement another oracle agent with additional ground truth 2D position of fallen objects (GT seg.+object+location) and achieve near-perfect results. The failure cases for this agent were mostly cases where the agent got caught in a narrow aisle and could not get out. These examples indicate that safe exploration in physical rooms is another challenge that must be overcome to fully solve this task. One potential solution for that is to learn a semantic search policy to set waypoints [35]. We leave this to future work.

Qualitative Results. In Figure 3, we have visualized example trajectories of the modular planning-based agent. The agent first estimates the relative distance and orientation of the fallen object and roughly what category this object might belong to. The planner then explores the environment and builds the map while navigating towards the goal. Once approaching the surrounding goal area, the visual signal dominates the agent's decisions. For example, after the agent identifies an object that seems consistent with what it hears, it will adjust its goal location, navigate towards it and execute the *found* action. This process intuitively resembles how humans might perform such a task. We will provide more demo videos in the supplementary materials.

4.5. Evaluation on Cross-Scene Generalization

Similar to other embodied AI challenges, we also hope to build autonomous navigation agents that generalize to unseen scenarios. To explicitly test agents' generalization abilities, we define two splits of the training set and test set for cross-scene evaluation. In the first split, the models are trained in the study rooms but tested in the kitchens. In the second split, we train models in the kitchens but test them in the study rooms. The results are reported in Table 2. All models have some degree of cross-scene performance drop. For example, the success rate of the modular planning approach decreased by 6%. We analyzed some failure cases and found that most of them come from the visual perception module. For example, the Mask-RCNN network sometimes fails to detect the fallen objects lying on a different color carpet, which requires agents to take additional time steps to find it.

5. Conclusions

We introduce a new task of embodied physical sound source localization in 3D virtual environments. To study this problem, we use the ThreeDWorld platform to synthesize a large-scale fallen object audio-visual dataset that includes 8000 instances of 30 physical objects dropping in 64 rooms. We implemented and evaluated several baseline agents as a first step to solving this challenge. Our experiments demonstrate that this task is challenging for current state-of-the-art methods. We further carried out a diagnostic analysis on modular planning-based agents to understand the factors that make this task challenging and to identify future directions for improving task performance. We believe this fallen object dataset will open up new opportunities to integrate physical reasoning abilities from multisensory data into embodied agents. We will release the code and dataset upon paper publication.

Future Works. The auditory perception model used by implemented embodied agents falls short of inferring physics from sound, one of the major bottlenecks for solving this new embodied AI challenge on finding fallen objects by integrating asynchronous audio-visual data. In future work, we are interested in learning object-centric neural implicit representations of impact sounds that could encode physical properties and transferring the models learned from this high-fidelity simulation to the real world.

Acknowledgement. This work was supported by MIT-IBM Watson AI Lab and its member company Nexplore, ONR MURI (N00014-13-1-0333), DARPA Machine Common Sense program, ONR (N00014-18-1-2847), NSF Grant BCS 1921501, and MERL.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *ECCV*, pages 208–224. Springer, 2020. 3

[2] Vinayak Agarwal, Maddie Cusimano, James Traer, and Josh H. McDermott. Object-based synthesis of scraping and rolling sounds based on non-linear physical constraints. *Digital Audio Effects (DAFx)*, pages 136–143, 2021. 1

[3] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 5

[4] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, pages 609–617, 2017. 3

[5] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, pages 435–451, 2018. 3

[6] Somil Bansal, Varun Tolani, Saurabh Gupta, Jitendra Malik, and Claire Tomlin. Combining optimal control and learning for visual navigation in novel environments. In *Conference on Robot Learning*, pages 420–429, 2020. 2

[7] Dhruv Batra, Angel X Chang, Sonia Chernova, Andrew J Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, et al. Rearrangement: A challenge for embodied ai. *arXiv preprint arXiv:2011.01975*, 2020. 2

[8] Matthew Chang, Arjun Gupta, and Saurabh Gupta. Semantic visual navigation by watching youtube videos. *NeurIPS*, 2020. 2, 3

[9] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[10] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 5, 7

[11] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *CVPR*, pages 15516–15525, 2021. 2, 3, 5, 6, 7

[12] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, pages 17–36, 2020. 1, 2, 3, 5

[13] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. *ICLR*, 2021. 3, 5

[14] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *NeurIPS*, 2021. 5, 7

[15] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. *CORL*, 2021. 3

[16] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. See, hear, explore: Curiosity via audio-visual association. *NeurIPS*, 2020. 3

[17] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM SIGGRAPH*, 2018. 3

[18] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *ECCV*, pages 758–775, 2020. 3

[19] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, et al. Threedworld: A platform for interactive multi-modal physical simulation. *NeurIPS*, 2021. 2, 3

[20] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, pages 9701–9707, 2020. 1, 2, 3, 5, 6, 7

[21] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *ICCV*, pages 7053–7062, 2019. 3

[22] Chuang Gan, Siyuan Zhou, Jeremy Schwartz, Seth Alter, Abhishek Bhandwaldar, Dan Gutfreund, Daniel LK Yamins, James J DiCarlo, Josh McDermott, Antonio Torralba, et al. The threedworld transport challenge: A visually guided task-and-motion planning benchmark for physically realistic embodied ai. *arXiv preprint arXiv:2103.14025*, 2021. 2

[23] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. *CORL*, 2021. 8

[24] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *European Conference on Computer Vision*, pages 658–676. Springer, 2020. 3

[25] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 3

[26] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *CVPR*, pages 324–333, 2019. 3

[27] Bruno L. Giordano and Stephen McAdams. Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *The Journal of the Acoustical Society of America*, 119:1171–1181, 2006. 1

[28] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *CVPR*, 2017. 2

[29] Meera Hahn, Devendra Chaplot, Shubham Tulsiani, Mustafa Mukadam, James M Rehg, and Abhinav Gupta. No rl, no simulation: Learning to navigate without navigating. *NeurIPS*, 2021. 2

[30] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 6

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[32] Elia Kaufmann, Mathias Gehrig, Philipp Foehn, René Ranftl, Alexey Dosovitskiy, Vladlen Koltun, and Davide Scaramuzza. Beauty and the beast: Optimal methods meet learning for drone racing. In *ICRA*, pages 690–696, 2019. 2

[33] Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 2

[34] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *CORL*, 2021. 2

[35] So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. Film: Following instructions in language with modular methods. *arXiv preprint arXiv:2110.07342*, 2021. 8

[36] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. In *CVPR*, pages 7539–7548, 2019. 3

[37] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, pages 631–648, 2018. 3

[38] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *CVPR*, pages 2405–2413, 2016. 3

[39] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816, 2016. 3

[40] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *CVPR*, pages 8494–8502, 2018. 2

[41] Senthil Purushwalkam, Sebastia Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. In *ICCV*, pages 1183–1192, 2021. 3

[42] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 5

[43] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP*, pages 4492–4496, 2020. 3

[44] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP*, pages 2357–2361, 2019. 3

[45] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. *ICCV*, 2019. 2

[46] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 5, 7

[47] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7574–7583, 2018. 3

[48] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, pages 10740–10749, 2020. 2

[49] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, C Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. *arXiv preprint arXiv:2108.03332*, 2021. 2

[50] Kun Su, Xiulong Liu, and Eli Shlizerman. Audeo: Audio generation for a silent performance video. *NeurIPS*, 2020. 3

[51] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, pages 436–454, 2020. 3

[52] James Traer, Maddie Cusimano, and Josh H. McDermott. A perceptually inspired generative model of rigid-body contact sounds. *Digital Audio Effects (DAFx)*, 2019. 1, 2, 3

[53] James A. Traer, Sam V. Norman-Haignere, and Josh H. McDermott. Causal inference in environmental sound recognition. *Cognition*, 214:104627, 2021. 1, 3

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5

[55] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. Visual room rearrangement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[56] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *ICLR*, 2020. 2

[57] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *CVPR*, pages 9068–9079, 2018. 2

[58] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *CVPR*, pages 11097–11107, 2020. 2

[59] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *CIRA*, pages 146–151, 1997. 5

[60] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, pages 570–586, 2018. 3

[61] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*, 2017. 2