



# Active Audio-Visual Separation of Dynamic Sound Sources

Sagnik Majumder<sup>1</sup>  and Kristen Grauman<sup>1,2</sup> 

<sup>1</sup> UT Austin, Austin, TX, USA

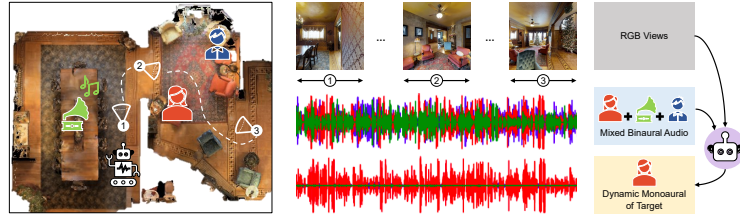
<sup>2</sup> Facebook AI Research, Austin, TX, USA  
{sagnik, grauman}@cs.utexas.edu

**Abstract.** We explore active audio-visual separation for dynamic sound sources, where an embodied agent moves intelligently in a 3D environment to *continuously* isolate the *time-varying* audio stream being emitted by an object of interest. The agent hears a mixed stream of multiple audio sources (e.g., multiple people conversing and a band playing music at a noisy party). Given a limited time budget, it needs to extract the target sound accurately at *every step* using egocentric audio-visual observations. We propose a reinforcement learning agent equipped with a novel transformer memory that learns motion policies to control its camera and microphone to recover the dynamic target audio, using self-attention to make high-quality estimates for current timesteps and also simultaneously improve its past estimates. Using highly realistic acoustic SoundSpaces [13] simulations in real-world scanned Matterport3D [11] environments, we show that our model is able to learn efficient behavior to carry out continuous separation of a dynamic audio target. Project: <https://vision.cs.utexas.edu/projects/active-av-dynamic-separation/>.

## 1 Introduction

Our daily lives are full of *dynamic audio-visual events*, and the activity and physical space around us affect how well we can perceive them. For example, an assistant trying to respond to his boss’s call in a noisy office might miss the initial part of her comment but then visually spot other noisy actors nearby and move to a quieter corner to hear better; to listen to a street musician playing a violin at a busy intersection, a pedestrian might need to veer around other onlookers and shift a few places to clearly hear .

These examples show how *smart sensor motion* is necessary for accurate audio-visual understanding of dynamic events. For audio sensing in an environment full of distractor sounds, variations in acoustic attributes like volume, pitch, and semantic content of a sound source—in concert with a listener’s proximity and direction from it—determine how well the listener is able to hear a target sound. However, just getting close or far is not always enough: effective hearing might require the listener to move around by *visually sensing* competing sound sources and surrounding obstacles, and discovering locations favorable to listening (e.g.,



**Fig. 1:** Active audio-visual separation of dynamic sources. Given multiple *dynamic* (time-varying, non-periodic) audio sources  $S$ , all mixed together, the proposed agent separates the audio signal by actively moving in the 3D environment on the basis of its egocentric audio-visual input, so that it is able to accurately retrieve the target signal at every step of its motion.

an intersection of two walls that reinforces listening through audio reflections, the front of a tall cabinet that can dull acoustic disturbances).

In this work, we investigate how to induce such intelligent behaviors in autonomous agents through audio-visual learning. Specifically, we propose the new task of *active audio-visual separation of dynamic sources*: given a stream of egocentric audio-visual observations, an agent must determine how to move in an environment with multiple sounding objects in order to *continuously* retrieve the *dynamic* (temporally changing) sounds being emitted by some object of interest, under the constraint of a limited time budget. See Figure 1. This task is relevant for augmented reality (AR) and mobile robotics applications, where a user equipped with an assistive hearing device or a service robot needs to better understand a sound source of interest in a busy environment.

Our task is distinct from related efforts in the literature. Whereas traditional audio-visual separation models extract sounds passively from pre-recorded videos [24,27,1,21,44,29,18,30,66,15,57,68,37], our task requires active placement of the agent’s camera and microphones over time. Whereas embodied audio-visual navigation [13,26,12,14,4] entails moving towards a sound source, our task requires recovering the sounds of a target object. The recent Move2Hear model performs active source separation [38] but is limited to *static* (i.e., periodic or constant) sound sources, such as a ringing phone or fire alarm—where recovering one timestep of the sound is sufficient. In contrast, our task involves *dynamic* audio sources and calls for extracting the target audio at *every step* of the agent’s movement. Variations in the observed audio arise not only from the room acoustics and audio spatialization, but also from the temporally-changing and fleeting nature of the target and distractor sounds. This means the agent must recover a new audio segment at every step of its motion, which it hears only once. The proposed task is thus both more realistic and more difficult than existing active audio-visual separation settings.

To address active dynamic audio-visual source separation, we introduce a reinforcement learning (RL) framework that trains an agent how to move to continuously listen to the dynamic target sound. Our agent receives a stream

of egocentric audio-visual observations in the form of RGB images and mixed binaural audio, along with the target category of interest (human voice, musical instrument, etc.) and decides its next action (translation or rotation of its camera and microphones) at every time step.

The technical design of our approach aims to meet the challenges outlined above. First, the proposed motion policy accumulates its observations with a recurrent network. Second, a transformer [59]-based acoustic memory leverages self-attention to simultaneously produce an estimate of the current segment of the dynamic target audio while refining estimates of past timesteps. The bi-directionality of the transformer’s attention mechanism helps capture regularities in acoustic attributes (e.g., volume, pitch, timbre) and semantic content (words in a speech, musical notes, etc.) and model their variations over a long temporal range. Third, the motion policy is trained using a reward that encourages movements conducive to the best-possible separation of the target audio at every step, thus forcing both the policy and the acoustic memory to learn about patterns in the acoustic and semantic characteristics of the target.

We validate our ideas with realistic audio-visual simulations from SoundSpaces [13] together with 53 real-world Matterport3D [11] environment scans, along with non-periodic sounds from many diverse human speakers, music, and other common background sources. Our agent successfully learns a motion policy for hearing its dynamic target more clearly in unseen scenes, outperforming the state-of-the-art Move2Hear [38] approach in addition to several baselines.

## 2 Related Work

*Audio(-Visual) Source Separation.* Substantial prior work uses audio alone to separate sounds in a passive (non-embodied) way, whether from single-channel (monaural) audio [54,56,61,35] or multi-channel audio that better surfaces spatial cues [41,65,20,19,62,67,28]. Bringing vision together with audio, approaches based on signal processing and matrix factorization [34,23,55,48,46,53,27], visual tracking [7,36], and deep learning [25,64,69,29] have all been explored, with particular recent emphasis on audio-visual separation of speech [1,24,21,44,2,18,30]. Other methods explore isolating a target audio source of interest [43,58,31,32,70].

Active models for hearing better include sound source localization methods in robotics that steer a microphone towards a localized source (e.g., [5,42,9]) and audio-visual methods that detect when people are speaking [3,60,6,?], which could help attend to one at a time. Most recently, Move2Hear [38] trains an embodied agent to move in a 3D environment using audio and vision in order to hear a target object. Although it is an important first step in tackling active separation, Move2Hear deals only with static sounds, as discussed above. This severely limits its real-world scope, where almost all sounds, including those of interest like human speech, music, etc., are time-varying. Addressing dynamic sounds is decidedly more complex: the model must leverage not only the agent’s surrounding geometry (to shut out undesirable interfering sounds) but also predict a continuously evolving audio sequence while hearing it just once within the mixed

input sound. Aside from substantially generalizing the task, compared to [38] our core technical contributions are a novel transformer [59]-based acoustic memory that does bi-directional self-attention to not only produce high-quality current separations on the basis of past separations but also refine past separations by taking useful cues from current separations, and a dense reward that promotes accurate dynamic separation at every step of the agent’s motion. We demonstrate our model’s advantages over [38] in results.

*Transformer Memory.* Transformers [59] have been shown to do very well on long-horizon embodied tasks like navigation and exploration [22,40,12,16,10,39,49]. The performance gains arise from a transformer’s ability to effectively leverage past experiences of the agent [22,40,39,10,49] and do cross-modal reasoning [12,16]. Different from these methods, our idea is to use a transformer as a memory model for capturing long-range acoustic correlations for audio-visual separation. Unlike prior approaches that show the successful use of transformers for passive separation [66,15,57,68,37], we train an agent to actively position itself in a 3D environment so that the transformer memory can improve current and past estimates of the target audio through bi-directional self-attention.

### 3 Task Formulation

We introduce a novel embodied task: active audio-visual separation of dynamic sound sources. In this task, an autonomous agent hears multiple *dynamic* audio sources of different types placed at various locations in an unmapped 3D environment, where the raw audio emitted by each source varies with time. The agent’s objective is to move around intelligently to isolate the *target source’s*<sup>3</sup> dynamic audio from the observed audio mixture at every step of its motion.

Our agent relies on both acoustic and visual cues to actively listen to the ever-changing target audio. The audio stream conveys cues about the spatial placement of the audio sources relative to the agent. Besides letting the agent see sounding objects within its field of view, the visual signal can help it avoid obstacles and go towards more closed spots in the 3D scene to suppress undesired acoustic interference. Both audio and vision can reveal how far the agent can venture in search of possibly favorable locations to cut out distractor sounds—without failing to follow the target. The temporal variations in the heard audio can also inform the agent about patterns in the fluctuations of the sound quality, intensity, and semantics (e.g., a voice raising and falling), letting it anticipate those attributes for future timesteps and move accordingly to maximize separation.

*Task Definition.* An episode of the proposed task begins by randomly instantiating multiple audio sources in a 3D environment by sampling their locations, types, and the dynamic audio tracks that play at the source positions. At every timestep, the agent receives a mixed binaural audio waveform, which depends on the surrounding scene’s geometric and material composition, the relative spatial

<sup>3</sup> e.g., a human speaker or instrument the agent wishes to listen to

arrangement of the agent and the sources, and the audio types (e.g., speech, musical instrument, etc.). One of the sources is the target. The agent’s goal is to extract the latent monaural signal playing at the target at every timestep.<sup>4</sup>

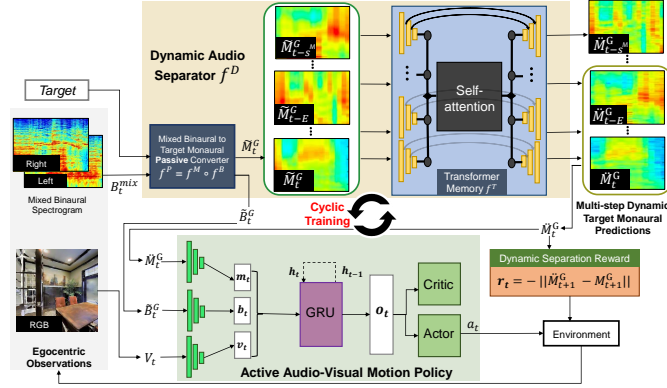
To succeed, the agent needs to leverage egocentric audio-visual cues to move intelligently so that it can predict the dynamic audio sequence. The agent cannot re-experience sounds from past timesteps, but it is free to revise its past predictions.

*Episode Specification.* We formally specify an episode by the tuple  $(\mathcal{E}, p_0, S_1(t), S_2(t), \dots, S_k(t), G^c)$ . Here,  $\mathcal{E}$  refers to the 3D scene in which the episode is executed,  $p_0 = (l_0, o_0)$  denotes the initial agent pose determined by its location  $l$  and orientation  $o$ , and  $S_i(t) = (S_i^w(t), S_i^l, S_i^c)$  specifies an audio source by its dynamic monaural waveform  $S_i^w$  as a function of the episode step  $t$ , location  $S_i^l$  and audio type  $S_i^c$ .  $k$  is the total number of audio sources of distinct types ( $S_1^c \neq S_2^c \neq \dots \neq S_k^c$ ) present in the scene, and  $G^c$  is the target audio type, such that  $G \in \{S_i\}$ . At each step, the agent’s objective is to listen to the binaural mixture of all sources and predict the dynamic signal  $G^w(t)$  associated with the target audio type  $G^c$ . Note that distinct human voices are considered distinct audio types. We limit the episode length to  $\mathcal{T}$  steps, thus assigning the agent a fixed time budget to retrieve the entire audio clip.

*Action Space.* At each timestep, the agent samples an action  $a_t$  from its action space  $\mathcal{A} = \{MoveForward, TurnLeft, TurnRight\}$  and moves on an unobserved *navigability graph* of the environment. While turns are always valid, *MoveForward* is not valid unless there is an edge from the current node to the next in the direction the agent is facing (i.e., no walls or obstacles exist there).

*3D Environment and Audio-Visual Simulation.* Following the state-of-the-art [13,12,38] in embodied audio-visual learning, we use the AI-Habitat simulator [52] with SoundSpaces [13] audio simulations and Matterport3D scenes [11]. Matterport3D provides dense 3D meshes and image scans of real-world houses and other indoor environments. SoundSpaces contains room impulse responses (RIR) to render realistic spatial audio at a resolution of 1 meter for Matterport3D. It models how sound waves from each source travel in the 3D environment and interact with the surrounding materials and geometry. In particular, the state-of-the-art RIRs simulate most real-world acoustic phenomena: direct sounds, early specular/diffuse reflections, reverberations, binaural spatialization, and frequency-dependent effects of materials and air absorption. See [13] for details. By using these highly perceptually realistic simulators together with real-world scanned environments and real-world audio data (cf. Sec. 5), we minimize the sim2real gap; we leave exploring transfer to a real robot for future work.

<sup>4</sup> The *monaural* source is the ground truth target, since it is devoid of all material and spatial effects due to the environment and the mutual positioning of the agent and the sources. Using a spatialized (e.g., binaural) audio as ground truth would permit undesirable shortcut solutions: the agent could move somewhere in the environment where the target is inaudible, and technically return the right answer of silence [38].



**Fig. 2:** Our model addresses active audio-visual separation of dynamic sources by leveraging a synergy between a *dynamic audio separator* and an *active audio-visual motion policy*. The dynamic separator  $f^D$  uses self-attention to continuously isolate a target signal  $M_t^G$  from its received mixed binaural  $B_t^{mix}$  on the basis of its past and current initial estimates  $\{\tilde{M}_{t-s}^G, \dots, \tilde{M}_t^G\}$ , while also using its current initial separation  $\tilde{M}_t^G$  to refine past final separations  $\{\tilde{M}_{t-E}^G, \dots, \tilde{M}_{t-1}^G\}$ . The motion policy uses the separator’s outputs,  $\tilde{M}_t^G$  and  $\tilde{B}_t^G$  and egocentric RGB images  $V_t$  to guide the agent to areas suitable for separating future dynamic targets.

In every episode, we place  $k$  point objects as the monaural sources in a Matterport 3D scene and render a binaural mixture  $B_t^{mix}$  of the audio coming to the agent from all the source locations, by convolving the monaural waveforms at each step with the corresponding RIRs and taking their mean. Our agent moves through the 3D spaces while receiving real-time egocentric visual and audio observations that are a function of the 3D scene and its surface materials, the agent’s current pose, and the sources’ positions.

## 4 Approach

We approach the problem with reinforcement learning. We train a motion policy to make sequential movement decisions on the basis of egocentric audio-visual observations, guided by dynamic audio separation quality. Our model has two main components (see Fig. 2): 1) an audio separator network and 2) an active audio-visual (AV) motion policy.

The separator network serves three functions at every step: 1) it passively separates the current target audio segment from its heard mixture, 2) it improves its past separations by exploiting their correlations in acoustic attributes like volume, pitch, quality, content semantics, etc. with the current separation, and 3) it uses the current separation to guide the motion policy towards locations favorable for accurate separation of future audio segments. The motion policy is trained to repeatedly maximize dynamic separation quality. It moves the agent in the 3D scene to get the best possible estimate of the *complete* target signal.

These two components share a symbiotic relationship, each feeding off of useful learning cues from the other while training. This allows the agent to learn the complex links between separation quality and the dynamic acoustic attributes of the target source, its location relative to the agent, the scene’s material layout (walls, floors, furniture, etc.), and the inferred spatial arrangement of distractor sources in the 3D environment.

#### 4.1 Dynamic Audio Separator Network

The dynamic audio separator network  $f^D$  receives the mixed binaural sound  $B_t^{mix}$  coming from all sources in the environment and the target audio type  $G^c$  at every step. Using these inputs, it produces estimates  $\ddot{M}^G$  of the monaural target at both the current and earlier  $E$  steps, i.e.,  $f^D(B_t^{mix}, G^c) = \{\ddot{M}_{t-E}^G, \dots, \ddot{M}_t^G\}$  (Fig. 2 top). We use short-time Fourier transform (STFT) to represent the monaural  $M$  and binaural  $B$  as audio magnitude spectrograms of dimensionality  $F \times N$  and  $2 \times F \times N$ , respectively, where  $F$  is the number of frequency levels,  $N$  is the number of overlapping temporal windows, and  $B$  has 2 channels (left and right).

The audio separation takes place by using two modules, such that  $f^D = f^P \circ f^T$ . First, the passive audio separator module  $f^P$  predicts the current monaural target  $\tilde{M}_t^G$ , given the audio goal category  $G^c$  and the binaural mixture  $B_t^{mix}$ . Next, the transformer memory  $f^T$  uses self-attention to combine cues from the current monaural prediction  $\tilde{M}_t^G$  and an external memory bank  $\mathcal{M}$  of  $f^P$ ’s past predictions to both produce an enhanced version  $\ddot{M}_t^G$  of the current monaural prediction and also refine the past  $E$  predictions  $\{\tilde{M}_{t-E}^G, \dots, \tilde{M}_{t-1}^G\}$ . Our multi-step prediction is motivated by the time-varying and fleeting nature of dynamic audio; we leverage semantic and acoustic structure in the target signal to both anticipate future targets and correct errors in past estimates. Next, we describe each of these modules in detail.

*Passive Audio Separator.* We adopt a factorized architecture [38] for the passive separator  $f^P$ , which comprises a binaural extractor  $f^B$  and a monaural predictor  $f^M$ . Given the mixed binaural and the target audio type,  $f^B$  predicts the target binaural  $\tilde{B}_t^G$  by estimating a real-valued ratio mask [29,64,69].  $f^M$  converts  $\tilde{B}_t^G$  into an initial estimate  $\tilde{M}_t^G$  of the current monaural target. Both  $f^B$  and  $f^M$  use U-Net like architectures [50] with ReLU activations. We enhance these initial monaural predictions with a transformer memory, as we will describe next.

*Transformer Memory.* The proposed memory is a transformer encoder [59] that encodes the current and past  $s^{\mathcal{M}}$  monaural predictions from  $f^M$ . Every transformer block has a pre-norm architecture that has been found to work particularly well for audio separation [57,68]. For storing past estimates from  $f^M$ , we maintain an external memory bank of size  $s^{\mathcal{M}}$ . At every step, we encode the past and current estimates using a multi-layer convolutional network to output a lower-dimensional monaural feature set  $\{e_{t-s^{\mathcal{M}}}, \dots, e_t\}$ . We feed these monaural features to the transformer along with sinusoidal positional encodings.

The transformer encoder uses the monaural features for its keys, values, and queries to build a joint representation through self-attention, such that all features across the temporal range  $[t - s^{\mathcal{M}}, t]$  can share information with each other. Further, the positional encoding for every feature informs the transformer about the feature’s temporal position in its context, thus helping it capture both short- and long-range similarities in the target dynamic signal.

On the transformer’s output side, we sample the feature set  $\{d_{t-E}, \dots, d_t\}$  associated with the timesteps in the range  $[t - E, t]$ , where  $E \leq s^{\mathcal{M}}$ , and decode them using another multi-layer convolution to obtain monaural estimates  $\{\tilde{M}_{t-E}^G, \dots, \tilde{M}_t^G\}$ . We also connect the convolutional encoder and decoder with additive skip connections for faster training and improved generalization [33].

This multi-step prediction is particularly useful in the dynamic setting, where at each step, the agent hears a new segment of the target sound. Consequently, the agent can benefit from an error-correction mechanism that uses the current estimates of the target to amend for mistakes in past estimates. Our results show that unlike RNN-based alternatives, the transformer memory naturally provides the agent with this option as it can exploit temporally local and global patterns in the dynamic audio—such as regularities in pitch and volume in human speech, and notes and timbre in music—in a bi-directional manner. By doing self-attention on an explicit storage of memory entries in a non-sequential fashion, it can accurately separate the current audio target and also refine past separations. The high quality of the current separations allows the agent to search for separation-friendly locations in its environment more extensively, by providing robustness in separation even when the agent temporarily passes through areas with high distractor interference.

Both modules  $f^P$  and  $f^T$  are trained in a supervised manner using the ground truth of the target binaural and monaural signals, as detailed in Sec. 4.3.

## 4.2 Active Audio-Visual Motion Policy

The audio-visual (AV) motion policy guides the agent in the 3D environment to predict the best possible estimate of the dynamic audio target at every timestep (Fig. 2 bottom). On the basis of real-time egocentric visual and audio inputs, the AV motion policy sequentially emits actions  $a_t$  that help continuously maximize the separation quality of the dynamic separator network  $f^D$  (Sec. 4.1). It has two modules: an observation encoder and a policy network.

*Observation Space and Encoding.* At each step, the AV motion policy receives the egocentric RGB image  $V_t$ , the current binaural separation  $\tilde{B}_t^G$  from  $f^B$ , and the current monaural prediction  $\tilde{M}_t^G$  from  $f^T$ .

The audio and visual streams provide the agent with useful complementary information for continuous separation of the dynamic target.  $\tilde{B}^G$  informs the agent about the relative spatial position (both distance and direction) of the target and also the scene’s materials and geometry, allowing it to anticipate how these factors might affect its future movements and separation quality. The binaural cue also prevents the agent from venturing too far away from the target, where



it would be unable to retrieve the current dynamic target, even with the help of its transformer memory. Besides informing the agent about the relation between separation quality and its pose relative to the target, as implicitly inferred from  $\tilde{B}^G$ ,  $\tilde{M}^G$  also allows the agent to track acoustic and semantic similarities in  $\ddot{M}^G$  over time by using its recurrent memory module (see below) and choose an action that could help maximize the next separation.

The visual signal  $V_t$  informs the policy about the 3D scene’s geometric configuration and helps it avoid collisions with obstacles. The synergy of vision and audio lets the agent learn about relations between the 3D scene’s geometry, semantics, and surface materials on the one hand, and the expected separation quality for different agent poses on the other hand.

We learn three separate CNN encoders to represent these inputs:  $v_t = \mathcal{F}^V(V_t)$ ,  $b_t = \mathcal{F}^B(\tilde{B}_t^G)$  and  $m_t = \mathcal{F}^M(\ddot{M}_t^G)$ . We concatenate  $v_t, b_t$  and  $m_t$  to obtain the complete audio-visual representation  $o_t$ .

*Policy Network.* The policy network in our model is a gated recurrent neural network (GRU) [17] followed by an actor-critic architecture. The GRU receives the current audio-visual representation  $o_t$  and its accumulated history of states from the last step  $h_{t-1}$  to produce an updated history  $h_t$  and also emit the current state feature  $s_t$ . The actor-critic module receives  $s_t$  and  $h_{t-1}$ , and predicts the policy distribution  $\pi_\theta(a_t|s_t, h_{t-1})$  and the current state’s value  $\mathcal{V}_\theta(s_t, h_{t-1})$ , where  $\theta$  are the policy parameters. The agent samples actions  $a_t$  from  $\mathcal{A}$  as per the policy distribution to interact with the 3D environment.

### 4.3 Training

*Dynamic Audio Separator Training.* As discussed in Sec 4.1, the separator network has two modules: the passive separator  $f^P$  that outputs  $\tilde{B}$  and  $\tilde{M}$ , and the transformer memory  $f^T$  that outputs  $\ddot{M}$ . We train both modules using the respective target spectrogram ground truths ( $B^G$  for binaural and  $M^G$  for monaural), which are provided by the simulator.

We train  $f^P$  with the passive separation loss:

$$\mathcal{L}_t^P = \|\tilde{B}_t^G - B_t^G\|_1 + \|\tilde{M}_t^G - M_t^G\|_1. \quad (1)$$

The training loss  $\mathcal{L}_t^T$  for the transformer memory aims to maximize the quality of both past and current monaural estimates,  $\{\ddot{M}_{t-E}^G, \dots, \ddot{M}_{t-1}^G\}$  and  $\ddot{M}_t^G$ , respectively:

$$\mathcal{L}_t^T = \frac{1}{E+1} \sum_{u=t-E}^t \|\ddot{M}_u^G - M_u^G\|_1. \quad (2)$$

Note  $f^P$  outputs step-wise predictions ( $\tilde{B}_t^G$  and  $\tilde{M}_t^G$ ), unlike  $f^T$  that uses past separation history to produce its separation output for the current step and its current output to correct past errors. Hence, we pretrain  $f^P$  using  $\mathcal{L}^P$  and a static dataset pre-collected from the training scenes by randomly instantiating the number of audio sources, their locations, the target audio type, the static

monaural segments for the sources, and the agent pose for each data point. Besides reducing the online computation cost by lowering the amount of trainable parameters, preliminary experiments showed that pretraining leads to stationarity in the distribution of  $\tilde{M}^G$  during the early phases of on-policy updates, thus stabilizing the training of the transformer.

Having pretrained  $f^P$ , we jointly train  $f^T$  and the AV motion policy with on-policy data while keeping the parameters of  $f^P$  frozen, since the agent’s trajectories directly affect the distribution of inputs for both  $f^T$  and the policy and hence, their training.

*Audio-Visual Motion Policy Training.* The dynamic separation task calls for a continuous retrieval of the time-varying target audio by the agent. Towards this goal, we train our motion policy with a dense RL reward:

$$r_t = -\|\ddot{M}_{t+1}^G - M_{t+1}^G\|_1. \quad (3)$$

At each step,  $r_t$  encourages the agent to act to maximize the next separation.

We train the policy using Decentralized Distributed PPO (DD-PPO) [63] with trajectory rollouts of 20 steps. The DD-PPO loss consists of a value loss, policy loss, and entropy loss to promote exploration (see Supp. Sec. 11).

*Joint Training.* We train the dynamic separator and AV motion policy by switching training between them after every parameter update.

## 5 Experiments

*Experimental Setup.* We instantiate each episode with  $k = 2$  audio sources that are placed randomly in the scene at least 8 m apart and specify one source as the target (Supp. details the dataset construction (Sec. 8) and shows results for  $k > 2$  (Sec. 6)). To clearly distinguish our task from audio-navigation, the agent is initially co-located with the target source at the episode start and is expected to fine-tune its position to isolate the dynamic audio target. We set the time budget  $\mathcal{T}$  to 20 for all episodes. We split 53 large Matterport3D scenes into non-overlapping train/val/test with 731K/100/1K episodes; we always evaluate the agent in unseen, unmapped environments.

Our dataset consists of 102 distinct types of sounds from three broad categories: speech, music, and background sounds. For speech, we use 100 different speakers from LibriSpeech [45]. For music, we combine multiple instruments from the MUSIC [69] dataset. For background sounds, we use non-speech and non-music sounds from ESC-50 [47], like running water, dog barking, etc. One of the 100 speakers (all speakers are considered distinct types) or music is the target. The distractors can be either background sounds, a speaker, or music. All sources play unique audio types. Our diverse dataset lets us evaluate different separation scenarios of varying difficulty: speech vs. speech, speech vs. music, or subtracting assorted background sounds.

There are 3,292 clips in total across all categories for use as monaural audio, each at least 20 s long. This helps avoid repetition in the monaural audio at each source in an episode with  $\mathcal{T} = 20$ . For unheard sounds, we split the clips into non-overlapping train/val/test in 18:3:4. For pretraining  $f^P$ , we generate 200K 1 s segments from the long clips and split it into train/val in 47:3, where the segments for each split are sampled from the corresponding split for long clips.

*Existing method and baselines.* We compare against the following methods:

- **Stand In-Place:** audio-only agent that does not change its pose throughout the episode, thereby separating the target in a completely passive way.
- **Rotate In-Place:** audio-only agent that rotates clockwise at its starting location at all steps to sample the mixed audio from all possible directions
- **DoA:** audio-only agent that takes one step away from the target, turns around, and samples the direct sound by facing its direction of arrival (DoA) [42] until the episode ends.
- **Random:** audio-only agent that randomly selects actions from the action space  $\mathcal{A}$  at all steps.
- **Proximity Prior:** an agent that selects random actions but stays inside a radius of 2 m (selected through validation) of the target so that it does not lose track of the dynamic goal audio. Unlike our agent, this agent assumes access to the privileged information of the ground truth distance to the target.
- **Novelty [8]:** a visual exploration agent trained with RL to maximize its coverage of novel locations in the environment within the episode budget. This helps sample diverse audio cues which can potentially aid dynamic separation.
- **Move2Hear [38]:** a state-of-the-art audio-visual RL agent for active source separation that was originally proposed to tackle static audio sources. We re-train this model on our task by using the dynamic separation reward (Sec. 4.3).

For fair comparison, we use the same dynamic audio separator  $f^D$  for all baselines and our agent, while the acoustic stream that the separator for each individual model receives depends on the its agent’s actions. While  $f^P$  shares its pre-trained parameters across all agents,  $f^T$  is trained separately for each agent by collecting on-policy data as per the individual agent’s policy. This helps disentangle the contributions to the separation quality stemming from the separator network and the policy designs, respectively. We set  $s^M$  to 19 and  $E$  to 14 on the basis of validation (see Supp. Sec. 4 for additional analysis on the choice of  $E$ ). These choices allow  $f^T$  to do accurate separation by leveraging the maximum amount of context possible in an episode ( $s^M = \mathcal{T} - 1$ ) while also refining past predictions. This way there is a balance between the amount of forward and backward information flow through  $f^T$ ’s bi-directional self-attention.

*Evaluation.* We evaluate all models in terms of mean separation quality of the dynamic monaural target over all  $\mathcal{T}$  steps, averaged over 1,000 test episodes and 3 random seeds. We use standard metrics: **STFT distance**, the error between separated and ground truth spectrograms, and **SI-SDR** [51], a scale-invariant measure of the amount of distortion present in the separated waveform.

**Table 1:** Active audio-visual dynamic source separation.

Model	<i>Heard</i>		<i>Unheard</i>	
	SI-SDR $\uparrow$	STFT $\downarrow$	SI-SDR $\uparrow$	STFT $\downarrow$
Stand In-Place	2.49	0.328	2.03	0.343
Rotate In-Place	2.50	0.327	2.04	0.343
DoA	2.78	0.313	1.88	0.342
Random	2.81	0.314	1.95	0.343
Proximity Prior	2.92	0.309	2.05	0.338
Novelty [8]	1.68	0.358	1.44	0.366
Move2Hear [38]	2.31	0.331	2.06	0.339
Ours	<b>3.93</b>	<b>0.273</b>	<b>2.57</b>	<b>0.318</b>
Ours w/o transformer memory $f^T$	2.32	0.330	2.02	0.340
Ours w/o multi-step predictions ( <i>i.e.</i> , $E = 0$ )	2.69	0.317	2.29	0.33
Ours w/o visual input $V_t$	3.61	0.284	2.40	0.324

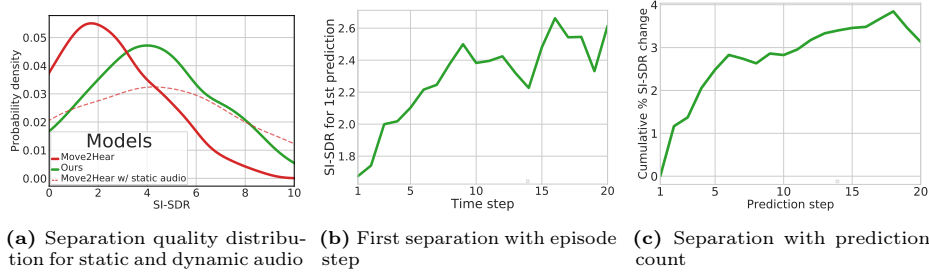
We provide more details about the data, spectrogram computation, baseline implementation, network architectures, training and evaluation in Supp Sec. 8-11.

### 5.1 Source Separation Results

Table 1 shows the separation quality of all models. The passive baselines Stand and Rotate In-Place fare worse than the ones that move and sample more diverse audio cues, like Random and Proximity Prior. However, despite being able to move, Novelty [8] performs poorly; in its effort to maximize coverage of the environment, it wanders too far from the target and fails to hear it in certain phases of its motion. DoA improves over the stationary baselines, since standing a step away from the target source allows it to sample a cleaner audio cue.

On the unheard sounds setting, Move2Hear [38] outperforms some baselines (e.g., DoA, Random, Novelty [8], etc.) but fares comparably against others (e.g., stationary baselines and Proximity Prior). This behavior can be attributed to the absence of the transformer memory  $f^T$  in Move2Hear. Although it is trained to maximize the separation quality at every step, its acoustic memory refiner is unable to handle the time-varying and transient nature of audio. To further illustrate our task complexity, we plot the distribution of separation performance (SI-SDR) of Move2Hear for both static and dynamic sources in Fig. 3a. The switch from dynamic to static audio results in a substantial degradation of separation quality (compare dotted and solid red curves: the mode of the SI-SDR distribution shifts leftward upon going from dynamic to static). This clearly shows the distinct new challenges posed by the dynamic source separation setting.

Our model outperforms all baselines and Move2Hear by a statistically significant margin ( $p \leq 0.05$ ). It overcomes the drop in Move2Hear’s separation quality induced by dynamic sources (compare solid red and green curves in Fig. 3a: the mode for dynamic audio shifts rightward upon using our method). Its performance demonstrates the impact of our active policy and long-term memory. Simply staying at or close to the source to be able to continuously hear the target (as done by Stand or Rotate In-Place, DoA, and Proximity Prior) is not enough. Our improvement over Move2Hear [38] further emphasizes the



**Fig. 3:** (a) Distribution of separation quality for static and dynamic audio (higher SI-SDR is better). (b) Quality at first separation as a function of time. (c) Separation quality for a target as a function of prediction count.

advantage of our transformer memory  $f^T$  in dealing with dynamic audio, both in terms of boosting separation when the agent is able to sample a cleaner signal, and providing robustness to the separator when the agent is passing through zones that are relatively less suitable for separation.

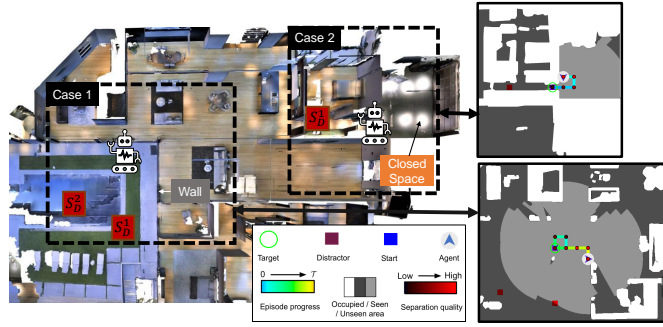
## 5.2 Model Analysis

*Ablations.* We report the ablation of our model components in Table 1 bottom. Our model suffers a severe drop in performance upon removing our transformer memory ( $f^T$ ). This shows that  $f^T$  plays a significant role in boosting dynamic separation quality.  $f^T$  learns joint representations of past and present separation estimates through self-attention, successfully leveraging long-range temporal relations in the audio signal for high-quality separation. The drop when  $E$  is set to 0 shows that the proposed multi-step predictions are also important in the dynamic setting, where the agent hears each segment of the time-varying signal only once and might not be able to separate a target very well at its first attempt. Our multi-step predictions let the agent improve past separations over time by extrapolating acoustic and semantic cues present in the current estimate. The visual cue  $V_t$  also contributes to our agent’s separation performance; without it, the agent is prone to collisions and has to solely rely on the separated binaural  $B_t^G$  to understand how the surrounding geometry and materials affect separation.

*Transformer Memory.* Next we analyze different aspects of our transformer memory  $f^T$  to understand how they boost our agent’s separation performance.

Fig. 3b shows how a longer memory leads to better initial estimates of the target audio segments. With more memory,  $f^T$  is better able to find consistencies in past estimates to improve the current estimate.

Fig. 3c shows how making more estimates of the same target using the multi-step prediction mechanism of  $f^T$  improves separation. Our model keeps refining all its predictions until late in the episode. This shows how  $f^T$  uses self-attention for enhancing past estimates on the basis of the current estimate through backward flow of useful information shared across multiple audio segments.



**Fig. 4:** Sample trajectories of our model. Our model not only uses audio-visual cues to find locations in the 3D scene that help dampen the distractors, but also utilizes its transformer memory to provide separation robustness and stability, while non-myopically looking for sweet spots with the potential to improve overall separation in the future.

See Supp. Sec. 5-7 for more analysis showing our model’s separation superiority with 1) microphone noise, 2) varying inter-source distances, and 3)  $k = 3$  sources.

*Qualitative Results.* In Fig. 4, we show two successful dynamic separation episodes of our method. In case 1, the agent starts outdoors and its motion choices are limited by the lack of potential acoustic barriers between the two distractors and the agent. Hence, it ventures away from the target in order to go indoors and make use of a wall to place itself in the acoustic shadow of the distractors, while still being able to sample the target audio. Such behavior is possible with our transformer memory, which maintains reasonable separation quality while the agent crosses regions of low separation quality looking for places that could improve future separation. In case 2, the agent makes short movements to find a closed space in its vicinity that not only blocks the distractor, but also reinforces its hearing of the target through acoustic reflections. See Supp. Sec. 2 and 3 for a discussion on our method’s failure cases and our model’s separation quality as a function of agent pose, respectively.

## 6 Conclusion

We introduced the task of active audio-visual separation of dynamic sources, where an agent must see and hear to move in an environment to continuously isolate the sounds of the time-varying source of interest. Our proposed approach outperforms the previous state-of-the-art in active audio-visual separation, as well as strong baselines and a popular exploration policy from prior work. In future work, we aim to extend our model to tackle sim2real transfer and moving sources, e.g., with audio source motion anticipation.

**Acknowledgements:** Thank you to Ziad Al-Halah for very valuable discussions. Thanks to Tushar Nagarajan, Kumar Ashutosh, and David Harwarth for feedback on paper drafts. UT Austin is supported in part by DARPA L2M, NSF CCRI, and the IFML NSF AI Institute. K.G. is paid as a research scientist by Meta.

## References

1. Afouras, T., Chung, J.S., Zisserman, A.: The conversation: Deep audio-visual speech enhancement. arXiv preprint arXiv:1804.04121 (2018)
2. Afouras, T., Chung, J.S., Zisserman, A.: My lips are concealed: Audio-visual speech enhancement through obstructions. arXiv preprint arXiv:1907.04975 (2019)
3. Alameda-Pineda, X., Horaud, R.: Vision-guided robot hearing. *The International Journal of Robotics Research* (2015)
4. Anonymous: Sound adversarial audio-visual navigation. In: Submitted to The Tenth International Conference on Learning Representations (2022), <https://openreview.net/forum?id=NkZq40EYN->, under review
5. Asano, F., Goto, M., Itou, K., Asoh, H.: Real-time sound source localization and separation system and its application to automatic speech recognition. In: *Eurospeech* (2001)
6. Ban, Y., Li, X., Alameda-Pineda, X., Girin, L., Horaud, R.: Accounting for room acoustics in audio-visual multi-speaker tracking. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018)
7. Barzelay, Z., Schechner, Y.Y.: Harmony in motion. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2007). <https://doi.org/10.1109/CVPR.2007.383344>
8. Bellemare, M.G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., Munos, R.: Unifying count-based exploration and intrinsic motivation. arXiv preprint arXiv:1606.01868 (2016)
9. Bustamante, G., Danes, P., Forgue, T., Podlubne, A., Manhès, J.: An information based feedback control for audio-motor binaural localization. *Autonomous Robots* (2018)
10. Campari, T., Eccher, P., Serafini, L., Ballan, L.: Exploiting scene-specific features for object goal navigation. In: *European Conference on Computer Vision*. pp. 406–421. Springer (2020)
11. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)* (2017), matterPort3D dataset license available at: [http://kaldir.vc.in.tum.de/matterport/MP\\_TOS.pdf](http://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf)
12. Chen, C., Al-Halah, Z., Grauman, K.: Semantic audio-visual navigation. In: *CVPR* (2021)
13. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: SoundSpaces: Audio-visual navigation in 3D environments. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer (2020)
14. Chen, C., Majumder, S., Al-Halah, Z., Gao, R., Ramakrishnan, S.K., Grauman, K.: Learning to set waypoints for audio-visual navigation. In: *International Conference on Learning Representations* (2021), <https://openreview.net/forum?id=cR91FAodFMe>
15. Chen, J., Mao, Q., Liu, D.: Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. arXiv preprint arXiv:2007.13975 (2020)
16. Chen, K., Chen, J.K., Chuang, J., Vázquez, M., Savarese, S.: Topological planning with transformers for vision-and-language navigation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11276–11286 (2021)

17. Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A.C., Bengio, Y.: A recurrent latent variable model for sequential data. In: *NeurIPS* (2015)
18. Chung, S.W., Choe, S., Chung, J.S., Kang, H.G.: Facefilter: Audio-visual speech separation using still images. *arXiv preprint arXiv:2005.07074* (2020)
19. Deleforge, A., Horaud, R.: The Cocktail Party Robot: Sound Source Separation and Localisation with an Active Binaural Head. In: *HRI 2012 - 7th ACM/IEEE International Conference on Human Robot Interaction*. pp. 431–438. ACM, Boston, United States (Mar 2012). <https://doi.org/10.1145/2157689.2157834>, <https://hal.inria.fr/hal-00768668>
20. Duong, N.Q., Vincent, E., Gribonval, R.: Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing* **18**(7), 1830–1840 (2010)
21. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619* (2018)
22. Fang, K., Toshev, A., Fei-Fei, L., Savarese, S.: Scene memory transformer for embodied agents in long-horizon tasks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 538–547 (2019)
23. Fisher III, J.W., Darrell, T., Freeman, W., Viola, P.: Learning joint statistical models for audio-visual fusion and segregation. In: Leen, T., Dietterich, T., Tresp, V. (eds.) *Advances in Neural Information Processing Systems*. vol. 13, pp. 772–778. MIT Press (2001), <https://proceedings.neurips.cc/paper/2000/file/11f524c3fbfeeca4aa916edcb6b6392e-Paper.pdf>
24. Gabbay, A., Shamir, A., Peleg, S.: Visual speech enhancement. *arXiv preprint arXiv:1711.08789* (2017)
25. Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A.: Music gesture for visual sound separation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10478–10487 (2020)
26. Gan, C., Zhang, Y., Wu, J., Gong, B., Tenenbaum, J.B.: Look, listen, and act: Towards audio-visual embodied navigation. In: *ICRA* (2020)
27. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 35–53 (2018)
28. Gao, R., Grauman, K.: 2.5d visual sound. In: *CVPR* (2019)
29. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3879–3888 (2019)
30. Gao, R., Grauman, K.: Visualvoice: Audio-visual speech separation with cross-modal consistency. *arXiv preprint arXiv:2101.03149* (2021)
31. Gu, R., Chen, L., Zhang, S., Zheng, J., Xu, Y., Yu, M., Su, D., Zou, Y., Yu, D.: Neural spatial filter: Target speaker speech separation assisted with directional information. In: Kubin, G., Kacic, Z. (eds.) *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*. pp. 4290–4294. ISCA (2019). <https://doi.org/10.21437/Interspeech.2019-2266>, <https://doi.org/10.21437/Interspeech.2019-2266>
32. Gu, R., Zou, Y.: Temporal-spatial neural filter: Direction informed end-to-end multi-channel target speech separation. *arXiv preprint arXiv:2001.00391* (2020)
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)



34. Hershey, J.R., Movellan, J.R.: Audio vision: Using audio-visual synchrony to locate sounds. In: *NeurIPS* (2000)
35. Huang, P., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Deep learning for monaural speech separation. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1562–1566 (2014). <https://doi.org/10.1109/ICASSP.2014.6853860>
36. Li, B., Dinesh, K., Duan, Z., Sharma, G.: See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 2906–2910 (2017). <https://doi.org/10.1109/ICASSP.2017.7952688>
37. Lu, W.T., Wang, J.C., Won, M., Choi, K., Song, X.: Spectnt: a time-frequency transformer for music audio. *arXiv preprint arXiv:2110.09127* (2021)
38. Majumder, S., Al-Halah, Z., Grauman, K.: Move2Hear: Active audio-visual source separation. In: *ICCV* (2021)
39. Mezghani, L., Sukhbaatar, S., Lavril, T., Maksymets, O., Batra, D., Bojanowski, P., Alahari, K.: Memory-augmented reinforcement learning for image-goal navigation. *arXiv preprint arXiv:2101.05181* (2021)
40. Mezghani, L., Sukhbaatar, S., Szlam, A., Joulin, A., Bojanowski, P.: Learning to visually navigate in photorealistic environments without any supervision. *arXiv preprint arXiv:2004.04954* (2020)
41. Nakadai, K., Hidai, K.i., Okuno, H.G., Kitano, H.: Real-time speaker localization and speech separation by audio-visual integration. In: *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*. vol. 1, pp. 1043–1049. IEEE (2002)
42. Nakadai, K., Lourens, T., Okuno, H.G., Kitano, H.: Active audition for humanoid. In: *AAAI* (2000)
43. Ochiai, T., Delcroix, M., Koizumi, Y., Ito, H., Kinoshita, K., Araki, S.: Listen to what you want: Neural network-based universal sound selector. In: Meng, H., Xu, B., Zheng, T.F. (eds.) *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*. pp. 1441–1445. ISCA (2020). <https://doi.org/10.21437/Interspeech.2020-2210>, <https://doi.org/10.21437/Interspeech.2020-2210>
44. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 631–648 (2018)
45. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: An asr corpus based on public domain audio books. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 5206–5210 (2015). <https://doi.org/10.1109/ICASSP.2015.7178964>
46. Parekh, S., Essid, S., Ozerov, A., Duong, N.Q.K., Pérez, P., Richard, G.: Motion informed audio source separation. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 6–10 (2017). <https://doi.org/10.1109/ICASSP.2017.7951787>
47. Piczak, K.J.: ESC: Dataset for Environmental Sound Classification. In: *Proceedings of the 23rd Annual ACM Conference on Multimedia*. pp. 1015–1018. ACM Press. <https://doi.org/10.1145/2733373.2806390>, <http://dl.acm.org/citation.cfm?doid=2733373.2806390>
48. Pu, J., Panagakis, Y., Petridis, S., Pantic, M.: Audio-visual object localization and separation using low-rank and sparsity. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 2901–2905. IEEE (2017)

49. Ramakrishnan, S.K., Nagarajan, T., Al-Halah, Z., Grauman, K.: Environment predictive coding for embodied agents. arXiv preprint arXiv:2102.02337 (2021)
50. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9351, pp. 234–241. Springer (2015), <http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a>, (available on arXiv:1505.04597 [cs.CV])
51. Roux, J.L., Wisdom, S., Erdogan, H., Hershey, J.R.: Sdr – half-baked or well done? In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 626–630 (2019). <https://doi.org/10.1109/ICASSP.2019.8683855>
52. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: ICCV (2019)
53. Sedighin, F., Babaie-Zadeh, M., Rivet, B., Jutten, C.: Two multimodal approaches for single microphone source separation. In: 2016 24th European Signal Processing Conference (EUSIPCO). pp. 110–114 (2016). <https://doi.org/10.1109/EUSIPCO.2016.7760220>
54. Smaragdis, P., Raj, B., Shashanka, M.: Supervised and semi-supervised separation of sounds from single-channel mixtures. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) Independent Component Analysis and Signal Separation. pp. 414–421. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
55. Smaragdis, P., Smaragdis, P.: Audio/visual independent components. In: in Proc. of International Symposium on Independent Component Analysis and Blind Source Separation (2003)
56. Spiertz, M., Gnann, V.: Source-filter based clustering for monaural blind source separation. In: in Proceedings of International Conference on Digital Audio Effects DAFx’09 (2009)
57. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J.: Attention is all you need in speech separation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 21–25. IEEE (2021)
58. Tzinis, E., Wisdom, S., Jansen, A., Hershey, S., Remez, T., Ellis, D.P., Hershey, J.R.: Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. arXiv preprint arXiv:2011.01143 (2020)
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
60. Viciano-Abad, R., Marfil, R., Perez-Lorenzo, J., Bandera, J., Romero-Garces, A., Reche-Lopez, P.: Audio-visual perception system for a humanoid robotic head. Sensors (2014)
61. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. On Audio, Speech and Lang. Processing (2007)
62. Weiss, R.J., Mandel, M.I., Ellis, D.P.: Source separation based on binaural cues and source model constraints. In: Ninth Annual Conference of the International Speech Communication Association, 2008 (2009)
63. Wijmans, E., Kadian, A., Morcos, A., Lee, S., Essa, I., Parikh, D., Savva, M., Batra, D.: Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. arXiv preprint arXiv:1911.00357 (2019)

64. Xu, X., Dai, B., Lin, D.: Recursive visual sound separation using minus-plus net. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 882–891 (2019)
65. Özgür Yilmaz, Rickard, S.: Blind separation of speech mixtures via time-frequency masking. In: IEEE TRANSACTIONS ON SIGNAL PROCESSING (2002) SUBMITTED (2004)
66. Zadeh, A., Ma, T., Poria, S., Morency, L.P.: Wildmix dataset and spectro-temporal transformer model for monaural audio source separation. arXiv preprint arXiv:1911.09783 (2019)
67. Zhang, X., Wang, D.: Deep learning based binaural speech separation in reverberant environments. IEEE/ACM transactions on audio, speech, and language processing **25**(5), 1075–1084 (2017)
68. Zhang, Z., He, B., Zhang, Z.: Transmask: A compact and fast speech separation model based on transformer. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5764–5768. IEEE (2021)
69. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European conference on computer vision (ECCV). pp. 570–586 (2018)
70. Žmolíková, K., Delcroix, M., Kinoshita, K., Ochiai, T., Nakatani, T., Burget, L., Černocký, J.: Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. IEEE Journal of Selected Topics in Signal Processing **13**(4), 800–814 (2019). <https://doi.org/10.1109/JSTSP.2019.2922820>