SoundAct: Learning Spatial Sound Awareness for Egocentric Robot Manipulation with Stereo Audio

Junsu Kim[1] and Kyungdon Joo[1]

***Abstract*— Humans naturally use auditory and visual cues to interact with objects beyond sight. However, most robot manipulation frameworks rely solely on vision, limiting their ability to handle audio-driven tasks (*e.g.*, ring-off) and out-of-view events. To address this, we propose `SoundAct`, a sound-aware egocentric robot manipulation framework that integrates stereo microphones with a wrist-mounted camera for beyond-sight spatial audio reasoning. We encode directional cues from stereo audio as magnitude spectrograms and fuse them with visual features via an attention mechanism, enabling the policy to adapt its reliance on auditory and visual inputs. We further introduce a spatial audio augmentation method to ensure robustness under audio distractors. We evaluate our method on a *beyond-sight ring-off* task and demonstrate effective manipulation of sound-source objects beyond sight. Video is available on `https://vision3d-lab.github.io/soundact/`**

I. Introduction

Humans jointly use spatial auditory awareness to interact with objects out of sight, such as turning off the ringing alarm clock hidden from the field of view (FOV). In contrast, many robot manipulation policies rely heavily on visual observations, often failing when events occur outside the FOV of the camera. Recent multimodal approaches have incorporated audio but typically treat it as a tactile signal [1], [2], [3] or as contextual information [4], rather than as a spatial cue for guiding action. Although sound source localization has been studied in robotics [5], [6], its integration into robot manipulation policy learning remains limited. Consequently, sound-based spatial awareness in robot manipulation is still underexplored.

To address this, we propose `SoundAct`, a spatial sound-aware egocentric robot manipulation framework that integrates stereo microphones with a wrist-mounted egocentric camera on a robotic arm. As illustrated in Fig. 1-(a), our system leverages stereo auditory cues to implicitly capture the direction of the sound source occurring beyond the FoV of the camera and incorporates them into policy learning, enabling the robot to respond to beyond-sight auditory events.

In this work, two key questions arise in enabling spatial sound-aware manipulation. First, how can stereo audio be represented to capture spatial cues for action reasoning? We address this by encoding stereo audio as the magnitude of spectrograms using the short-time Fourier transform (STFT), enabling the policy to implicitly infer the direction of sound sources beyond the visual field. Second, how can the policy robustly focus on the target audio under distractors? We introduce the spatial sound augmentation method that independently scales noise audio applied to the left and right channels. This channel-specific scaling enables the policy to remain focused on the target audio, even in the presence of louder audio distractors.

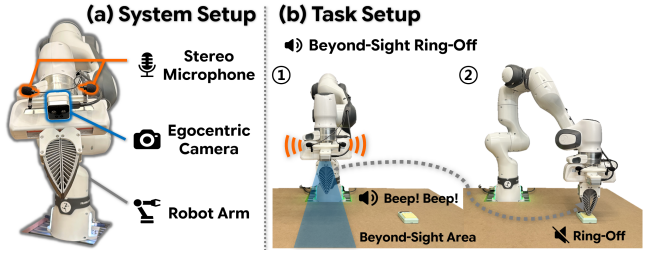

Fig. 1. **Overview of `SoundAct`.** (a) System setup with a stereo microphone array and a wrist-mounted egocentric camera on a robot arm. (b) Task setup for spatial sound-aware manipulation, where the robot localizes a ringing alarm clock outside its field of view using stereo audio and turns it off once it becomes visible.

To evaluate the spatial sound awareness of our framework, we introduce a *beyond-sight ring-off* task, in which the robot first moves toward the ringing alarm clock outside its visual FoV using stereo audio and then turns it off once the object becomes visible, as illustrated in Fig. 1-(b).

II. Method

A. Egocentric Stereo-Visual Perception

Hardware setup. Our hardware platform includes a Franka Research 3 robot arm integrated with a wrist-mounted Intel RealSense D405 camera. For auditory perception, two omnidirectional Maono AU-XLR10 condenser microphones are mounted laterally relative to the viewing direction of the camera, capturing stereo audio via a Behringer UMC404HD audio interface.

Data collection setup. Raw sensory streams, including wrist-view video (60 Hz), stereo audio (48 kHz), and robot states (end-effector pose at 100 Hz and gripper state at 30 Hz), are synchronized and resampled to a unified rate of 20 Hz. During training, the policy uses a 2-step observation of images and end-effector poses, together with 2-second stereo audio resampled to 16 kHz.

B. Stereo Audio-Visual Policy Architecture

As depicted in Fig. 2, we design a multi-modal policy that integrates visual, auditory, and proprioceptive observations to enable manipulation both within and beyond sight.

[1]Junsu Kim and Kyungdon Joo are with the Artificial Intelligence Graduate School, UNIST, Ulsan, South Korea. {`joonsu0109`, `kyungdon`}`@unist.ac.kr`

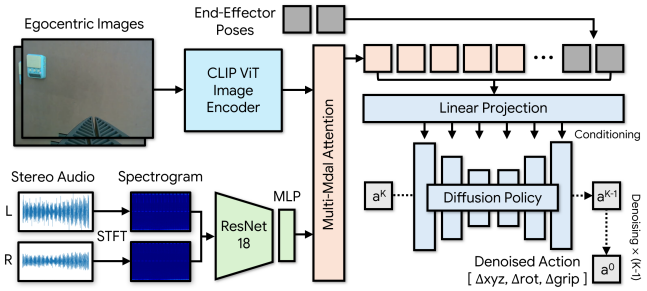

Fig. 2. **SoundAct model architecture.**

Vision encoding. Egocentric images are processed using a CLIP-pretrained ViT encoder [7]. The [CLS] tokens are concatenated to form vision latent features, following [1].

Stereo audio encoding. The stereo waveform is converted into spectrograms via the STFT. We take the magnitude of the spectrograms and process them using a ResNet-18 encoder. The resulting features are then projected through an MLP to align their dimensions with the visual features.

Feature fusion. Visual and audio latent features are concatenated and processed by the Transformer with self-attention. This allows the policy to dynamically weigh visual and auditory cues based on target visibility. The fused representation is then concatenated with the end-effector pose and used as a condition for action prediction.

Action prediction. Robot actions are predicted via the mechanism of diffusion policy [8], where a 1D convolutional UNet iteratively denoises action sequences. Conditioned on the fused multimodal representation, the policy predicts 16-step trajectories through a 50-step denoising process. The policy is trained with an MSE loss to predict relative trajectories.

Spatial audio augmentation. To enable the policy to distinguish target audio from distractors and avoid overfitting to absolute magnitude, we introduce a spatial audio augmentation method. We sample noise from ESC-50 [9], following ManiWAV [1], and apply independent random scaling to the left and right channels to simulate audio distractors.

III. Experiments

A. Task Setup

We evaluate our framework on a *beyond-sight ring-off* task, where the robot infers the direction of an alarm clock outside its initial FOV and navigates toward it using stereo audio, turning it off once visible. For training, we collect 60 teleoperated episodes with the target uniformly placed within a 20 cm × 20 cm rectangular region on either the left or right beyond-sight area. For evaluation, we conduct 10 trials in the in-distribution setting and 5 trials in audio distractor settings, including music, conversation, typing, another beep, and another clock alarm.

B. Performance Comparison

Fig. 3 and Table I present the qualitative results and quantitative comparisons, respectively.

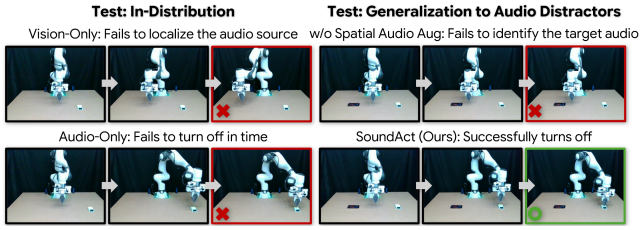

Fig. 3. **Beyond-sight ring-off evaluation.** The robot first moves toward a ringing alarm clock outside its field of view using stereo audio, then turns it off once it becomes visible. A third-person camera is used only for visualization.

TABLE I

SUCCESS RATES ON THE *beyond-sight ring-off* TASK.

Method	Success / Trials
Task: Beyond-Sight Ring-Off (In-Distribution)	
(a) Modalities	
Vision-Only	5/10
Audio-Only	7/10
Mono Audio (Only Left)	8/10
Stereo Audio (Ours)	**10/10**
(b) Audio Representation	
Log-Mel Spectrogram	5/10
STFT Spectrogram 4-Ch	2/10
STFT Spectrogram 2-Ch (Ours)	**10/10**
Task: Beyond-Sight Ring-Off (Audio Distractors)	
(c) Audio Generalization	
No Audio Augmentation	0/5
Background Audio Augmentation	0/5
Spatial Audio Augmentation (Ours)	4/5

Modality comparison. As shown in Table I-(a), vision-only misses out-of-view cues, while mono audio lacks directional information. Audio-only lacks manipulation precision. In contrast, our framework achieves the best performance by effectively integrating spatial audio with visual feedback.

Audio representation. Table I-(b) shows log-mel spectrograms lose spatial cues, while 4-channel stereo (magnitude and phase) causes excessive sensitivity. Using 2-channel stereo magnitude yields the best localization performance.

Audio generalization. Table I-(c) demonstrates the effectiveness of spatial audio augmentation. The policy shows improved robustness to distractor loudness and context variations. However, it struggles with similar periodic beep patterns to the target.

IV. Conclusion

We propose SoundAct, a spatial sound-aware egocentric robot manipulation framework that leverages stereo microphones and a wrist-mounted camera to address beyond-sight manipulation challenges. SoundAct enables reliable directional guidance of sound sources using stereo audio, even in the presence of audio distractors. However, the current evaluation is limited to a task-specific environment. Therefore, future work will focus on evaluating the policy in in-the-wild settings and extending it to operate within more complex visual environments for robust object interaction.

REFERENCES

[1] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song, "ManiWAV: Learning robot manipulation from in-the-wild audio-visual data," in *8th Annual Conference on Robot Learning*, 2024.

[2] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu, "See, hear, and feel: Smart sensory fusion for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1368–1378.

[3] M. Du, O. Y. Lee, S. Nair, and C. Finn, "Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning," in *Robotics: Science and Systems XVIII, New York City, NY, USA, June 27 - July 1, 2022*, K. Hauser, D. A. Shell, and S. Huang, Eds., 2022.

[4] R. Wang, H. Geng, T. Li, P. Wu, F. Wang, G. Anumanchipalli, T. Darrell, B. Li, P. Abbeel, J. Malik, and A. A. Efros, "The sound of simulation: Learning multimodal sim-to-real robot policies with generative audio," in *9th Annual Conference on Robot Learning*, 2025.

[5] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 2017.

[6] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *European conference on computer vision*. Springer, 2020, pp. 17–36.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[8] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[9] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.