Hear-to-Act: Spatial Audio-Guided Robot Control on Humanoid and Manipulator Platforms

Zhiyang Dou[1], Yuxin Ray Song[1], Yuxiang Ma[1], Weirui Ye[1], Pengfei Ye[1], Shuyang Xu[2], Jason Xinyu Liu[1], Kaichen Zhou[1], Chuang Gan[3], Taku Komura[2], Paul Pu Liang[1], Wojciech Matusik[1]
[1]Massachusetts Institute of Technology [2]The University of Hong Kong [3]University of Massachusetts Amherst

***Abstract*—The direction of a sound is a perceptual cue that robots seldom exploit, although it reaches them through occlusions and outside the field of view. Humans hear this direction with two ears: the inter-aural time and level differences between them let us localize a source and orient toward it, so binaural audio carries enough spatial information to drive a spatially responsive reaction. Turning that binaural hearing into *action on real robots* requires three pieces most robots lack: robot-mounted spatial hearing, a way to collect spatial audio-conditioned robot demonstrations, and controllers that act on the perceived sound. We take a first step toward all three. We equip two platforms (a Unitree G1 humanoid and an OpenManipulator-X arm) with a binaural microphone pair at left/right "ear" positions, reproducing the inter-aural cues that human listeners rely on. To collect responsive demonstrations, we build a teleoperation rig that streams the robot's binaural hearing to the operator's two ears in real time, so the operator perceives the scene spatially *as the robot does* and teleoperates the matching reaction. Finally, we make explicit what each platform contributes today. On the humanoid, the contribution is data collection: each teleoperated episode yields paired spatial-audio demonstrations. On the manipulator, the operator directly teleoperates the end-effector to collect synchronized binaural audio, video, and motion, and these episodes train an initial sound-conditioned controller for end-effector orientation. Together these take a first step from binaural perception toward robot action: robot-frame data collection on the humanoid, and manually collected manipulator demonstrations for a learned sound-conditioned controller.**

I. Introduction

Sound is one of the few signals that reaches a robot from outside its field of view and through occlusions. A spoken request, a falling object, or a hazard signal conveys to an embodied agent both *what* is occurring and *where* it originates. Humans exploit this constantly: binaural hearing lets us localize a source and orient toward (or away from) it before we ever see it. Equipping robots with this capability would make them more responsive and more capable collaborators in human environments.

Most audio-driven motion models map *what* is heard (speech, music, or a sound class) to motion, discarding the spatial layout of the scene. A recent line of work departs from this by conditioning motion generation on *binaural* audio [15], showing that the resulting motions are both semantically appropriate and spatially responsive, turning and reaching toward or away from the source. In effect, that binaural audio is a perceptual representation that already encodes the spatial structure a policy needs to act. That work, however, produces *simulated human* motion. Carrying it to *action* on physical robots raises three practical questions a motion generator does not answer:

- **Perception.** How should a robot *hear* spatially, with sensing that matches the binaural cues these models consume?
- **Data.** How do we collect demonstrations that pair a robot's spatial hearing with the appropriate physical reaction?
- **Action.** How can robot-frame sound demonstrations train a controller, and what simple analytical rule checks that the binaural cue is usable?

This paper contributes a first answer to each. (i) We mount a binaural microphone pair at left/right ear positions on two embodiments, a Unitree G1 humanoid and an OpenManipulator-X arm [11] (Sec. III). (ii) We build a teleoperation pipeline that streams the robot's binaural audio to the operator's two ears, letting the operator hear the scene as the robot does and teleoperate the response, yielding paired audio–motion robot demonstrations (Sec. IV). (iii) We separate what each platform demonstrates today: on the humanoid the working contribution is the data-collection rig itself, while on the manipulator we collect data by direct end-effector teleoperation and use those demonstrations to train an initial sound-conditioned controller. We retain a proportional inter-aural level controller as an analytical baseline for the binaural signal (Sec. VI). Together they trace a path from binaural perception to robot action.

II. Related Work

Audio-driven motion. Mapping audio to motion is well studied for music and dance [14, 12, 7, 13], but these methods condition on the *content* of a signal and discard the spatial layout of the scene. A recent line of motion-generation work closes this gap by conditioning on binaural audio [15]; we carry that spatial representation onto real robots and toward robot action. **Spatial audio and acoustic perception.** Binaural source localization from a two-channel ear signal, optionally aided by head rotation, is long standing [3], and acoustic simulation and learned acoustic fields [1, 6] bring spatial sound into embodied learning. Audio paired with vision supports source separation and 3D spatial-sound modeling [17, 16, 4]. We take the simplest spatial cue, inter-aural level, directly as a control input on hardware. **Hearing to action.** Closing the loop from a perceptual signal to action is the crux of deploying spatial audio on robots. We instantiate it for an unusual modality: a binaural signal that already encodes *where*, driving an orienting action and, through teleoperated demonstrations, learnable reactions. Robot audition itself is established: binaural microphones, inter-aural features, and stereo localization are standard [3]. Our

contribution is not a new feature or encoder but the combination that turns spatial audio into usable robot data: robot-mounted ears, an operator-grounded teleoperation interface that self-labels reactions in the robot's own acoustic frame, and the dataset this produces; the learned policy is a proposed consumer of that data rather than a claimed result.

III. Binaural Hearing on Robots

To let a robot perceive sound the way such spatial-audio models record their input, we give each platform a *binaural* pair of microphones, one "left ear" and one "right ear," rather than a single microphone or a flat array. Two ear-spaced microphones reproduce the inter-aural time and level differences that underlie human and learned binaural localization, and they keep the sensing modality identical to the binaural signals used to drive motion, so perception and downstream control share one representation. Binaural localization from exactly this kind of two-channel signal, optionally aided by head rotation, is well established [3].

Figure 1 shows the two installations. On the humanoid (Fig. 1a), the microphones sit at the left and right sides of the head, giving an ear baseline close to a human's. On the manipulator (Fig. 1b), the pair is mounted near the wrist so that the end-effector behaves like a small movable "head": rotating the last joints changes the inter-aural geometry, which, as in human listening, disambiguates front/back and sharpens localization. Both setups use only two commodity microphones, keeping the sensor cheap and directly compatible with binaural audio processing.

Why a binaural pair. We use a binaural pair here and leave larger microphone arrays to future work. A pair is the minimal sensor that produces the inter-aural time and level differences binaural perception relies on; it is low-cost, two microphones and a stereo stream; and it mirrors human hearing, placing two ears on the robot so it senses space the way a listener does. It also keeps the robot's input in the *same* two-channel format that binaural-conditioned motion models consume, so perception and policy share one representation. Mounting the pair on a movable part also turns motion into an active sensing cue: a small turn that changes which ear is louder supplies the head-rotation signal that resolves front/back, so acting and perceiving draw on the same degree of freedom.

Setup. The humanoid is a Unitree G1 carrying a binaural headset whose two microphones sit at the left and right sides of the head; the OpenManipulator-X [11] carries the same pair near its wrist, together with a camera so it records video alongside audio. For teleoperation the operator wears a head-mounted display and two hand controllers that command the robot's motion, plus earphones that play the robot's left and right channels (Fig. 2). The robot streams its video and binaural audio to the operator and receives motion commands in return, so a single rig captures synchronized video, audio, and motion with no extra instrumentation of the scene. Table I contrasts the two embodiments, and the command mapping is embodiment-specific. On the humanoid the operator's head and hand motions retarget to the G1's neck and arms, so turning toward a sound maps naturally onto the robot's own head rotation. On the manipulator the operator directly commands the end-effector, and the last-joint yaw plays the role of the "head" that turns the ears, a mapping that is less anthropomorphic but drives the sensing degree of freedom directly. Both mappings are grounded in the same percept, the robot's own binaural stream, so the recorded reactions stay comparable across embodiments despite the different control interfaces.

TABLE I: **Two embodiments for binaural hearing.** Both carry the same two-microphone "ear" pair and are taught by teleoperation in which the operator hears the robot's own binaural stream; they differ in scale, in what physically turns toward the source, and in what each demonstration records.

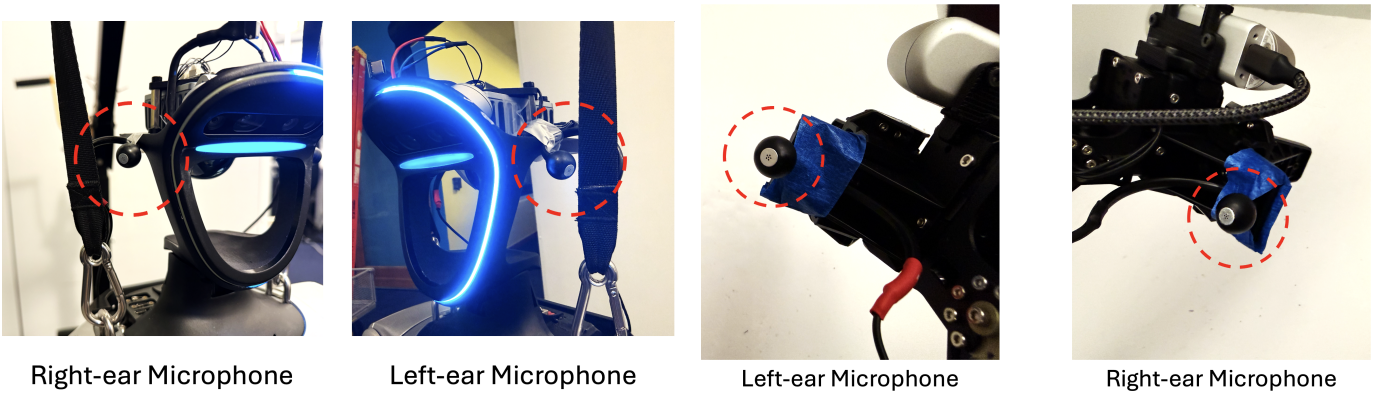
	Unitree G1	**OpenManipulator-X**
Ear placement	sides of the head	near the wrist
Inter-aural baseline	$\approx$ human head width	end-effector scale
"Head" that orients	neck and whole body	last joints (yaw)
Teleoperation	whole-body (HMD + controllers)	end-effector
Recorded streams	binaural audio, motion	binaural audio, video, motion
Proposed policy output	whole-body target	end-effector yaw
Demonstrated here	teleoperated data collection	manual teleop and controller training
Analytical baseline	not run closed loop	proportional ILD orienting rule

IV. Teleoperated Spatial-Audio Data Collection

Learning sound-responsive behavior needs demonstrations that pair what the robot hears with how it should move. Collecting these is awkward with conventional teleoperation: the operator stands outside the robot's acoustic vantage point and cannot perceive the scene the way the robot's ears do. We close this gap by routing the robot's hearing to the operator.

As shown in Fig. 2, the left and right microphone channels captured on the robot are played back live to the operator's left and right ears. The operator therefore experiences the same inter-aural cues (direction, distance, and intensity) as the robot, and teleoperates the reaction it should produce: turning toward a call, recoiling from a sudden bang, or reaching in the direction of an object's sound. Recording the robot's binaural audio together with the commanded motion gives paired demonstrations, the robot analogue of the human spatial audio–motion data such models are trained on. The same pipeline serves both embodiments: whole-body teleoperation for the humanoid and end-effector teleoperation for the manipulator. Because the operator's perception is grounded in the robot's own sensors, the collected reactions are spatially consistent with what a learned policy would later receive at deployment.

In practice the two microphone channels are streamed from the robot to the operator's left and right earphones alongside the robot's video, so the operator simultaneously sees and *hears* from the robot's vantage point while commanding motion. The pair is carried as a low-latency two-channel stream over the same link as the teleoperation, kept time-aligned with the



(a) Unitree G1 humanoid head.

(b) OpenManipulator-X arm.

Fig. 1: **Binaural hearing on two robots.** We mount a left/right microphone pair at ear-like positions on (a) a Unitree G1 humanoid head and (b) an OpenManipulator-X arm near the end-effector. The ear-spaced baseline preserves the inter-aural cues used for binaural localization and keeps the sensing modality identical to the spatial-audio representation that drives motion.

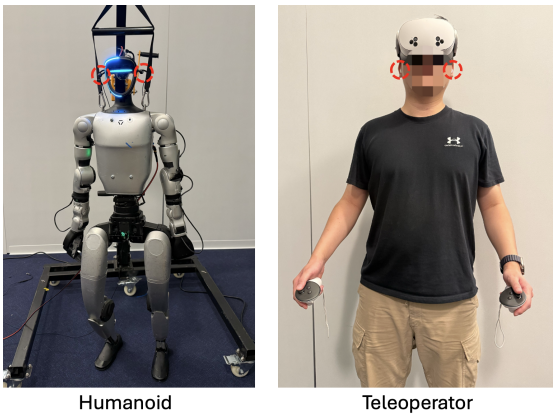


Fig. 2: **Teleoperated spatial-audio data collection.** The robot's binaural microphones (left) are streamed in real time to the operator's two ears (right). The operator, wearing a headset for whole-body or end-effector teleoperation, hears the scene spatially *as the robot does* and performs the appropriate reaction, yielding paired (binaural audio, robot motion) demonstrations.

video so the heard event and the seen event arrive together. Because the operator reacts to what they hear, end-to-end audio latency should stay within the tens of milliseconds over which audiovisual events are still perceived as synchronous; we have not yet adopted a formal perceptual threshold from the literature or measured the rig's latency, and the exact transport, codec, and buffering that bound it are part of hardening the rig for larger collections. On the manipulator each manually teleoperated demonstration logs the robot's synchronized *video*, *binaural audio*, and the commanded trajectory. These episodes train the sound-conditioned end-effector controller described in Sec. VI; the time-locked video is retained for a later audio-visual policy rather than used as evidence for the present audio controller. Streaming from the robot's own microphones rather than from the room keeps the recorded reactions aligned with the signal a deployed policy receives, and it adds only a pair of earphones to an existing teleoperation rig, so the collection inherits the scalability of the underlying setup.

The humanoid as a data-collection platform. On the Unitree G1 our present role is collection rather than autonomous control. The operator, wearing the head-mounted display and hand controllers, hears the robot's left and right channels through matched earphones and commands a whole-body reaction, so each episode records the humanoid's binaural audio together with the whole-body motion the operator produced. This is the humanoid analogue of the human spatial audio–motion data that binaural motion generators are trained on, captured directly in the robot's own acoustic frame. We do not yet run a closed-loop controller on the humanoid; the recorded demonstrations are the substrate on which the policy of Sec. VI is meant to be trained. No autonomous humanoid controller is claimed in this version.

What a demonstration captures. Each episode is a short interaction: a sound is produced somewhere around the robot, the operator, hearing it through the robot's ears, turns the robot toward or away from it, and the synchronized binaural audio and commanded motion are logged as a single training pair. Varying the source position, the sound type, and the robot's starting pose across episodes builds a set that spans the headings and acoustic conditions a deployed policy must handle. Because the operator's percept is the robot's own stereo stream, every recorded reaction is already expressed in the frame the policy will see at deployment, so no relabeling or spatial calibration between operator and robot is required.

V. Spatial Audio as a Robot Data Modality

An underused modality. The data that drives robot learning is overwhelmingly visual and proprioceptive. Large cross-embodiment corpora pair camera frames with joint and end-effector state [10], and where audio appears at all it is treated as a single-channel contact cue [2] or folded into an audio-visual stream for manipulation [8]. The *directional* content of sound, the inter-aural differences that encode *where* a source is rather than *what* it is, is rarely recorded. Yet this is the part of the signal that reaches a robot through occlusion and from behind, and the part a binaural pair is built to capture. Spatial audio therefore belongs alongside vision and proprioception in

TABLE II: **Dataset card for the teleoperated spatial-audio record.** Every episode is self-labeled by the operator's commanded reaction and expressed in the robot's own acoustic frame, so the record needs no microphone-array calibration and no source-position annotation.

Field	Content
Modalities	binaural audio (two channels), motion; plus video on the manipulator
Motion target	whole-body target (Unitree G1) or end-effector yaw (OpenManipulator-X)
Label source	the operator's commanded reaction, recorded as the demonstration target
Reference frame	the robot's own binaural (acoustic) frame
Coverage axes	source azimuth and range, sound type, room and reverberation, starting pose
Real and simulated	teleoperated episodes, extensible with rendered binaural audio

robot datasets. The obstacle has been collection, not usefulness: there is no standard way to log a robot's hearing together with the reaction it should produce, and a recorded waveform with no grounding in the robot's frame is hard to supervise. The teleoperation rig of Sec. IV addresses this. Because the operator hears through the robot's own ears, the reaction is already expressed in the robot's acoustic frame, and the binaural stream the operator acts on is the stream a policy receives at deployment. The supervision is operator-grounded: each demonstration carries its own spatial label in the form of the commanded motion, so the pipeline needs no microphone-array calibration, no source-position annotation, and no post-hoc relabeling.

A synchronized, scalable record. Each episode logs binaural audio, video, and motion on one clock, so the three modalities are aligned frame by frame rather than registered after the fact. That alignment is what makes the record a training pair: the inter-aural cues at time t sit next to the motion the operator produced in response, the structure behavior cloning consumes in Eq. (11). The initial manipulator controller is trained from the audio and state stream, while the time-locked video is retained for an audio-visual policy left to future work, so the same demonstrations support a richer model without recollection. Coverage scales along four axes: source azimuth and range, sound type from speech to transient impacts, room geometry and reverberation, and the robot's starting pose, since the mounting geometry that produces the inter-aural cues moves with the body. Sweeping these on hardware is slow, and acoustic simulation [1, 6] offers a complementary route: rendering binaural responses for many source positions and reverberation conditions populates the axes that are tedious to instrument physically, and pairing simulated audio with the same binaural feature front end of Sec. VI gives a way to study sim-to-real transfer for spatial acoustics before committing operator time. We treat the teleoperated record and its simulated extension as one dataset that grows along these axes, with real demonstrations anchoring the reactions and simulation broadening the acoustic conditions under which they are seen.

We summarize the record as a dataset card in Table II. Two properties make it usable as robot training data rather than raw logs: every episode is expressed in the robot's own acoustic frame, and its label is the operator's commanded reaction, so supervision comes for free with collection. The same card describes both embodiments; only the motion-target field and the presence of video differ between the humanoid and the manipulator.

VI. Proposed Architecture

We propose a policy π_θ that maps a short window of the robot's binaural audio, with its proprioceptive state, to a motion command that orients toward an active source,

$$\mathbf{a}_t = \pi_\theta(\mathbf{B}_t,\, \mathbf{s}_t), \tag{1}$$

where $\mathbf{B}_t$ is the binaural feature window defined below, $\mathbf{s}_t \in \mathbb{R}^{n_s}$ is the proprioceptive state (joint angles and end-effector pose), and $\mathbf{a}_t$ is the command, an end-effector yaw on the manipulator or a whole-body target on the humanoid.

Binaural feature extraction. For each channel $c \in \{L, R\}$ we take the short-time Fourier transform $X_c[k,\tau]$ and an F-band log-mel spectrogram with mel filterbank $\mathbf{M}$ [9],

$$S_c[m,\tau] = \log\Big(\textstyle\sum_k \mathbf{M}[m,k]\,|X_c[k,\tau]|^2 + \varepsilon\Big), \tag{2}$$

Azimuth is carried by the inter-aural level and phase differences, taken per mel band,

$$\Lambda[m,\tau] = S_R[m,\tau] - S_L[m,\tau], \tag{3}$$

$$\phi[m,\tau] = \angle\Big(\textstyle\sum_k \mathbf{M}[m,k]\, X_R[k,\tau]\,\overline{X_L[k,\tau]}\Big), \tag{4}$$

together with a broadband inter-aural level difference from the per-channel RMS energy over a W-sample window,

$$E_c = \sqrt{\tfrac{1}{W}\textstyle\sum_n x_c[n]^2}, \qquad \Delta_\tau = \frac{E_R - E_L}{E_R + E_L + \varepsilon} \in [-1, 1]. \tag{5}$$

Stacking the per-frame features (column τ)

$$\mathbf{z}_\tau = \big[\, S_L,\, S_R,\, \Lambda,\, \cos\phi,\, \sin\phi,\, \Delta \,\big]_\tau \in \mathbb{R}^{n_f}, \quad n_f = 5F{+}1, \tag{6}$$

over an L-frame window gives the policy input $\mathbf{B}_t = [\mathbf{z}_{t-L+1}, \ldots, \mathbf{z}_t]^\top \in \mathbb{R}^{L \times n_f}$.

Transformer encoder. We project the window to width d and prepend a proprioceptive token $\mathbf{h}_s = \mathbf{W}_s \mathbf{s}_t$,

$$\mathbf{H}^{(0)} = \big[\, \mathbf{h}_s;\ \mathbf{B}_t\mathbf{W}_{\text{in}} + \mathbf{P} \,\big] \in \mathbb{R}^{(L+1)\times d}, \tag{7}$$

with $\mathbf{W}_{\text{in}} \in \mathbb{R}^{n_f \times d}$ and sinusoidal positional encoding $\mathbf{P}$. We stack N Transformer encoder layers with gated residuals,

$$\mathbf{Z}^{(l)} = \mathrm{LN}\big(\mathbf{H}^{(l-1)} + \gamma_a^{(l)}\,\mathrm{MHA}(\mathbf{H}^{(l-1)})\big), \tag{8}$$

$$\mathbf{H}^{(l)} = \mathrm{LN}\big(\mathbf{Z}^{(l)} + \gamma_f^{(l)}\,\mathrm{FFN}(\mathbf{Z}^{(l)})\big), \tag{9}$$

where MHA is multi-head self-attention built from scaled dot-product attention $\mathrm{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\big(\mathbf{Q}\mathbf{K}^\top/\sqrt{d_k}\big)\mathbf{V}$, $\mathrm{FFN}(\mathbf{u}) = \mathrm{SiLU}(\mathbf{u}\mathbf{W}_1)\mathbf{W}_2$, and $\gamma_a^{(l)}, \gamma_f^{(l)} \in \mathbb{R}$ are learned skip gates. Self-attention over the window lets the policy integrate evidence as the source and the robot's own rotation change the inter-aural cues, the active-sensing effect that helps resolve front/back [3]. We pool the final layer to $\mathbf{h} = \mathrm{Pool}(\mathbf{H}^{(N)}) \in \mathbb{R}^d$.

Policy heads and training. A lightweight head maps $\mathbf{h}$ to the motion command, and an auxiliary head regresses the source azimuth $\hat{\theta}_t$ as an interpretable intermediate,

$$\mathbf{a}_t = \mathrm{MLP}_a(\mathbf{h}), \qquad \hat{\theta}_t = \mathrm{MLP}_\theta(\mathbf{h}). \tag{10}$$

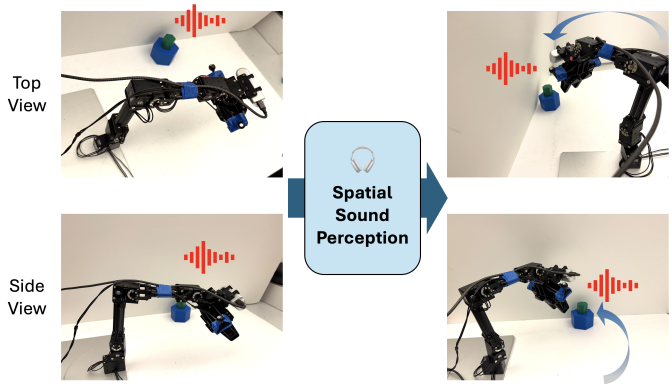


Fig. 3: **Analytical intensity baseline.** Top and side views of the manipulator orienting toward an active sound source: a proportional controller on the inter-aural level difference Δ (Eq. (12)) rotates the end-effector "head" until the source is centered.

The policy is trained by behavior cloning on the teleoperated $(\mathbf{B}_t, \mathbf{s}_t, \mathbf{a}_t^\star)$ demonstrations of Sec. IV, with the operator's commanded motion $\mathbf{a}_t^\star$ as the target,

$$\mathcal{L} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left\| \mathbf{a}_t - \mathbf{a}_t^\star \right\|_2^2 + \frac{\lambda}{|\mathcal{T}_\theta|} \sum_{t \in \mathcal{T}_\theta} \mathrm{wrap}_\pi\big(\hat{\theta}_t - \theta_t^\star\big)^2, \quad (11)$$

where $\mathcal{T}_\theta \subseteq \mathcal{T}$ are the frames that carry an azimuth label and $\mathrm{wrap}_\pi(x) = ((x + \pi) \bmod 2\pi) - \pi$ handles angular wraparound. Because the operator heard exactly the robot's binaural stream, training and deployment use the same sensor frame. In this submission the manipulator controller is trained from directly teleoperated end-effector episodes, while the humanoid contribution remains the collection system and its logged whole-body demonstrations. A deployed streaming version of Eq. (1) would encode each new audio window and issue a command to the low-level controller before the next window arrives; closed-loop validation of that controller follows in the next hardware study.

Manipulator controller and analytical baseline. The manipulator data are collected by manual end-effector teleoperation: the operator hears the OpenManipulator-X binaural stream, drives the wrist yaw or end-effector pose by hand, and supplies the target command $\mathbf{a}_t^\star$ used in Eq. (11). To make the spatial cue auditable, we also evaluate a non-learned baseline on the same binaural mounting. This baseline reads out only the broadband inter-aural level difference Δ of Eq. (5), written Δ_t at control step t, through a proportional controller,

$$\dot{\psi}_t = K_p \, \Delta_t, \quad (12)$$

rotating the end-effector "head" toward the louder ear ($K_p > 0$, with ψ measured toward that ear). The closed loop drives $\Delta_t \to 0$; of the two equal-energy configurations, facing the source is the stable equilibrium and the rear is unstable, reflecting the front/back ambiguity of a two-microphone pair.

Figure 3 shows this intensity-based baseline from top and side views: from an arbitrary starting pose, the manipulator rotates until the source is centered, and it re-orients when the source is moved. Across repeated trials with a single dominant source it settles facing the source, validating the binaural mounting of Sec. III and the streaming pipeline of Sec. IV. The reaction is paced by the analysis window: a longer window steadies the estimate of Δ_t but slows the response. The learned controller is meant to extend beyond this one-dimensional loudness readout by using the full binaural feature window and the teleoperated motion targets. The baseline also marks the limit of a purely analytical stereo cue: it cannot resolve front from back, weigh competing sources, or decide whether to approach or retreat by sound category.

Limitations. The present setup is minimal, and its limits are clear. Two microphones leave a front/back ambiguity that a static pose cannot resolve; reverberation, background noise, and the robot's own motor noise corrupt the inter-aural cues; and the current policy orients in azimuth but does not yet interpret *what* it hears. The demonstrations so far are small-scale and assume a single dominant source. This version also uses *audio alone*: it does not yet fuse vision, which would help resolve front/back, reject distractor sounds, and ground the source spatially. Scaling the data and fusing complementary modalities should narrow these gaps. The manipulator controller is trained from direct teleoperation, and closed-loop evaluation across rooms, source types, and multiple simultaneous sources is a clear next step. During collection the operator receives both the robot's video and its binaural audio, so a demonstrated reaction may draw on vision as well as sound; isolating the audio-only contribution, for instance by comparing how accurately an operator localizes a source under audio-only versus audio-plus-video conditions, is left to a controlled study.

VII. Discussion and Future Work

We presented a pipeline that carries spatial-audio responsiveness from generated human motion toward real robots: binaural hearing on a humanoid and an OpenManipulator-X arm, teleoperation that lets an operator hear as the robot hears and record reactions, manipulator demonstrations collected by direct manual control, an initial sound-conditioned controller trained from those episodes, and a proportional inter-aural baseline that exposes the limits of a one-cue rule. Several directions follow. A learned front/back model using spectral (HRTF) cues would remove the two-microphone ambiguity. Conditioning the policy on sound *category*, grounded in large-scale audio corpora [4], would let a single policy approach a call yet retreat from an alarm. Two evaluation directions would sharpen the claim: moving the task beyond "orient toward sound" to behaviorally meaningful decisions, such as approaching a person who calls for help, backing away from an alarm, or reaching toward a sounding object, where the right action depends on *what* is heard and not only *where* it is; and measuring the value of the modality directly, by comparing audio-guided behavior against vision-only and proprioception-only baselines and by analyzing failure modes such as wrong turns, delayed responses, and front/back ambiguity resolved through active sensing. Executing the predicted motion through a whole-body controller [5] would extend the arm's orienting gesture to a full humanoid reaction, and we plan to move the manipulator results from the OpenManipulator-X to higher-degree-of-freedom arms such as the Franka. Fusing the recorded video into a full

Vision-Language-Action policy is the larger goal. Acoustic simulation [1] and learned binaural localization [3] offer routes to scale the data and close the sim-to-real gap for spatial acoustics. Binaural audio is a low-cost, task-relevant signal for connecting perception to action, and the pipeline we describe is one route from binaural-conditioned motion generation toward robots that perceive and respond to sound on hardware.

REFERENCES

[1] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems*, 35:8896–8911, 2022.

[2] Maximilian Du, Olivia Lee, Suraj Nair, and Chelsea Finn. Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning. In *Robotics: Science and Systems (RSS)*, 2022.

[3] Guillermo García-Barrios, Daniel Aleksander Krause, Archontis Politis, Annamaria Mesaros, Juana M Gutiérrez-Arriola, and Rubén Fraile. Binaural source localization using deep learning and head rotation information. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 36–40. IEEE, 2022.

[4] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[5] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, et al. ASAP: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *arXiv preprint arXiv:2502.01143*, 2025.

[6] Zitong Lan, Yiduo Hao, and Mingmin Zhao. Resounding acoustic fields with reciprocity, 2025. URL https://arxiv.org/abs/2510.20602.

[7] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10234–10243, 2023.

[8] Zeyi Liu, Cheng Chi, Eric Cousineau, Naveen Kuppuswamy, Benjamin Burchfiel, and Shuran Song. ManiWAV: Learning robot manipulation from in-the-wild audio-visual data. In *Conference on Robot Learning (CoRL)*, 2024.

[9] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. *SciPy*, 2015:18–24, 2015.

[10] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024.

[11] ROBOTIS Co., Ltd. OpenMANIPULATOR-X. https://emanual.robotis.com/docs/en/platform/openmanipulator_x/overview/, 2017. Accessed: 2026-06-03.

[12] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022.

[13] Taoran Tang, Jia Jia, and Hanyang Mao. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1598–1606, 2018.

[14] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023.

[15] Shuyang Xu, Zhiyang Dou, Mingyi Shi, Liang Pan, Leo Ho, Jingbo Wang, Yuan Liu, Cheng Lin, Yuexin Ma, Wenping Wang, and Taku Komura. MOSPA: Human motion generation driven by spatial audio. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[16] Xudong Xu, Dejan Markovic, Jacob Sandakly, Todd Keebler, Steven Krenn, and Alexander Richard. Sounding bodies: modeling 3d spatial sound of humans using body pose and audio. *Advances in Neural Information Processing Systems*, 36:44740–44752, 2023.

[17] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 587–604, 2018.