

Audio Embodied Intelligence: Auditory Perception, Reasoning, and Interaction for Multimodal Agents

Wentao Lei*, Tianxin Xie*, Kai Jiang*, and Li Liu†

Abstract—Auditory learning is becoming increasingly important for embodied intelligence because sound provides information that may be unavailable or ambiguous in vision and language alone, including out-of-view events, physical interactions, spatial structure, task intent, and human affective states. Current research is progressing from isolated auditory perception toward embodied agents that use sound to support reasoning, action, and interaction. However, relevant studies remain fragmented across different tasks, system settings, and research communities, making it difficult to identify their shared structure, compare methodological progress, and evaluate how auditory information contributes to embodied behavior. This work addresses these issues through a two-axis loop–stage taxonomy. The loop axis comprises the Physical Interaction Loop, the Semantic Grounding-and-Action Loop, and the Socio-Affective Interaction Loop, while the stage axis characterizes how auditory information contributes to Percept, Reason, and Interact. Within this taxonomy, we examine how existing approaches acquire and represent auditory information, integrate it with other modalities, and use it to support embodied perception, reasoning, action, and interaction. We also summarize the associated methods, datasets, benchmarks, and evaluation metrics. Our work reveals a broad shift from task-specific auditory modules toward more integrated embodied systems, while showing that advances in auditory perception have not yet been consistently translated into reliable behavioral improvement. Major challenges therefore lie in robust auditory grounding under real-world conditions, effective integration of sound into decision-making and control, and evaluation that can determine the contribution of audition to embodied behavior.

Index Terms—Embodied intelligence, robot audition, audio-visual navigation, speech instruction, vision-language-action models, social robots, affective interaction.



1 INTRODUCTION

Embodied intelligence refers to an agent’s capacity to perceive, reason, and act through continuous interaction with the physical and social world. Its defining characteristic is the coupling between perception and action: sensory observations inform the agent’s internal state and behavior, while the consequences of action generate new observations that support subsequent reasoning and adaptation. Embodiment therefore involves more than applying perceptual models to robot data; intelligence must be situated in an agent, grounded in an environment, and expressed through action or interaction.

Research in embodied AI has progressed from task-specific navigation and manipulation toward multimodal agents that connect perception, language, planning, and control. Interactive simulation and robot learning have enabled the systematic study of situated behavior, while foundation models and vision–language–action models have expanded the range of instructions, environments, and skills that a single system can address [1], [2], [3], [4], [5]. Despite this progress, most embodied systems remain centered on vision, language, and action, leaving other sources of sensorimotor evidence comparatively underexplored.

Audition provides a complementary channel that is temporally fine-grained, non-contact, and available when relevant objects or events are occluded or outside the camera’s field of view. Impact and contact sounds reveal manipulation dynamics, spatial audio provides cues about source locations and acoustic scene structure, speech conveys task intent, and prosodic or other vocal cues provide evidence about user state. These properties make sound a natural modality for connecting perception, reasoning, action, and interaction within an embodied loop.

Figure 1 places this work in the context of recent publication growth. The OpenAlex counts show a marked increase in work indexed under embodied AI and embodied intelligence after 2020, together with growth in the corresponding audio-related subsets. Rather than establishing the maturity of individual tasks, the trend indicates that a previously dispersed body of auditory embodied research has become large enough to warrant a focused synthesis.

The central difficulty is that auditory learning for embodied intelligence does not correspond to a single task, but spans a family of problems with different inputs, inferred states, and behavioral consequences. A microphone array for sound source localization, a binaural navigation policy, a spoken-instruction robot, a contact-audio manipulation model, and an empathy-aware social robot all use sound, but they contribute to different stages and roles within situated perception–action or perception–interaction processes. Some works provide an auditory capability used by an embodied system, whereas others explicitly use auditory feedback to adapt subsequent action or interaction and thereby implement a closed loop. If these works are grouped only by modality, task name, or neural architecture, the logic of embodiment is obscured. We therefore organize the field by asking three questions. First, *what does the agent listen for?* Second, *what state, meaning, or user condition does it infer?* Third, *how does this inference inform physical action, task execution, communicative interaction, or their evaluation?*

In this work, we use *embodied audio* to denote auditory input or output that is explicitly connected to a situated agent, environment, task, or action-relevant evaluation, rather than sound that is merely recognized or generated as an isolated offline category. Under this view, audio plays three complementary roles. First, it serves as *physical evidence*: impact sounds, echoes, motion noise, and contact acoustics reveal hidden events, material properties, spatial layout, and action consequences. Second, it serves as *semantic and paralinguistic evidence*:

• Wentao Lei, Tianxin Xie, Kai Jiang, and Li Liu are with The Hong Kong University of Science and Technology (Guangzhou).
• *Equal contribution. †Corresponding author: Li Liu, avrillliu@hkust-gz.edu.cn.

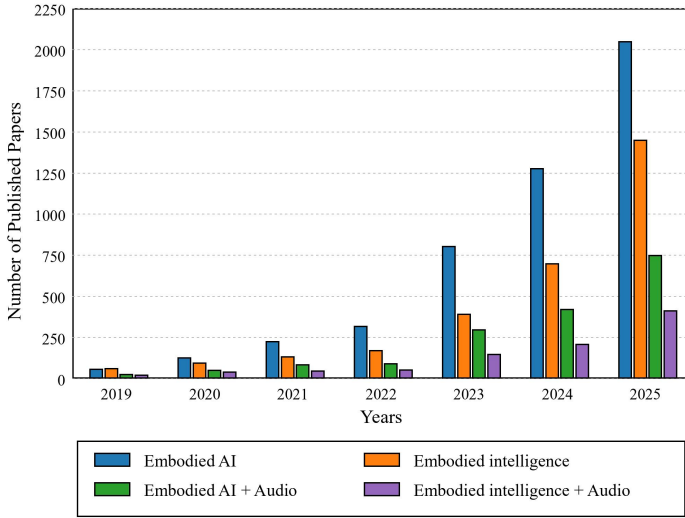


Fig. 1. Growth in publications on OpenAlex for key topics related to auditory-centered embodied intelligence: “Embodied AI,” “Embodied intelligence,” “Embodied AI + Audio,” and “Embodied intelligence + Audio” (2019–2025). Audio-related subsets are deduplicated by OpenAlex work ID across the terms audio, auditory, acoustic, sound, and speech.

speech conveys task intent and constraints, while prosody, rhythm, pauses, hesitation, laughter, sighs, intensity, and tone provide cues about how an instruction is expressed and whether the user may be uncertain, urgent, engaged, or in need of repair. Third, audio can be an *interaction signal*: spoken responses, nonverbal sounds, sonification, and spatialized feedback can shape action timing, turn-taking, safety warnings, clarification, trust calibration, and social adaptation. The key distinction is therefore not whether a system contains an audio encoder or completes an entire closed loop, but whether auditory information makes an explicit, situated contribution to embodied perception, reasoning, action, interaction, or evaluation.

Existing surveys provide substantial coverage of several related research traditions, including robot audition and acoustic localization, audio-visual perception and navigation, spoken human–robot interaction, affective and social robotics, and embodied AI. However, these traditions are generally reviewed within their respective task or disciplinary boundaries. What remains less developed is a shared framework for explaining how physical, semantic, and socio-affective forms of auditory information contribute to situated perception, reasoning, action, and social interaction.

This perspective motivates a two-axis loop–stage taxonomy. The loop axis distinguishes three coupled forms of auditory embodiment: the *Physical Interaction Loop*, in which sound provides evidence about the environment and the agent’s actions; the *Semantic Grounding-and-Action Loop*, in which speech, sound, and language convey task-relevant meaning; and the *Socio-Affective Interaction Loop*, in which paralinguistic and expressive cues inform human-centered adaptation. The stage axis distinguishes *Percept*, *Reason*, and *Interact*, corresponding respectively to the extraction of situated cues, the construction of task-relevant state or meaning, and the use of that state to guide action or communication. Crossing the two axes yields nine loop–stage categories, summarized in Table 1 and formally defined in Section 2.

The taxonomy characterizes the embodied role of auditory information and the stage to which a work primarily contributes. It does not require every included system to implement a complete closed loop, nor does it assume that a paper belongs exclusively to one category.

What makes audio embodied?

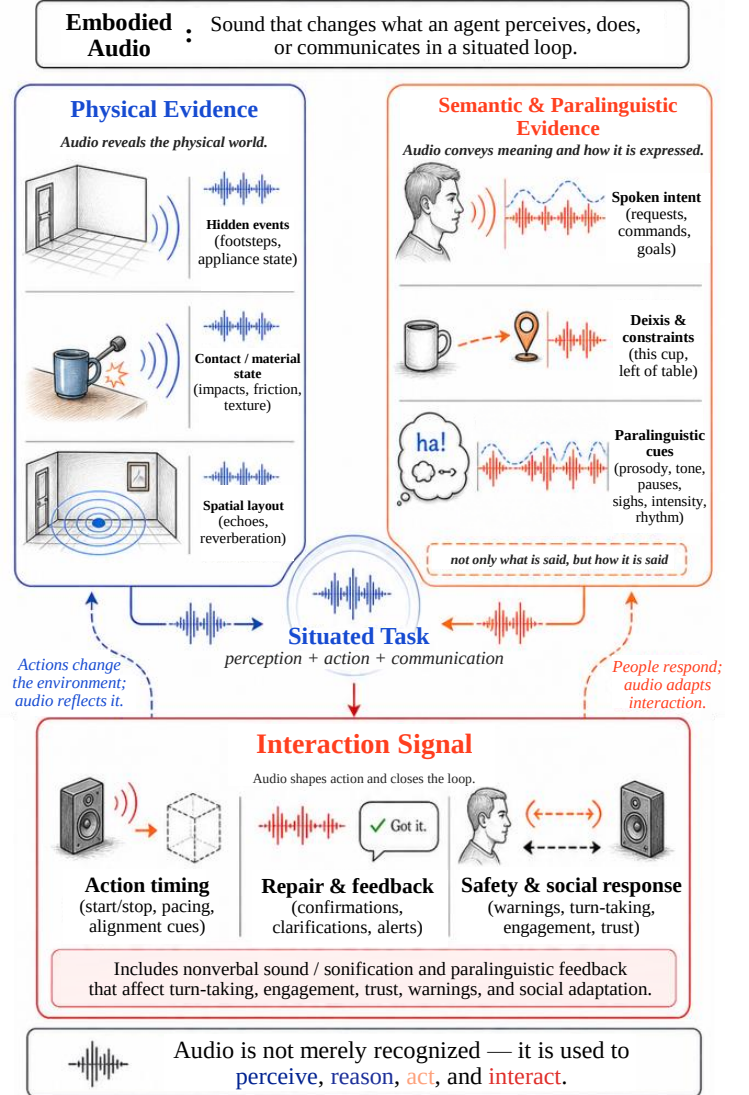


Fig. 2. Definition of embodied audio and its roles in an embodied loop. Audio can serve as physical evidence, semantic and paralinguistic evidence, and an interaction signal that changes belief, policy, feedback, action, and social response.

This work makes the following four contributions:

- We introduce a two-axis loop–stage taxonomy for auditory-centered multimodal learning in embodied intelligence, comprising three embodied loops and three functional stages.
- We provide a structured review of representative methods and systems across the nine loop–stage categories, showing how auditory information supports embodied perception, grounding, action, and interaction.
- We summarize task families, benchmarks, and evaluation metrics for assessing whether audio improves embodied behavior, rather than only improving standalone perception accuracy.
- We discuss key challenges and future directions for developing robust, real-time, and socially aligned audio-aware embodied agents.

The remainder of the work is organized as follows: Section 2 formalizes the taxonomy and the embodied auditory learning problem;

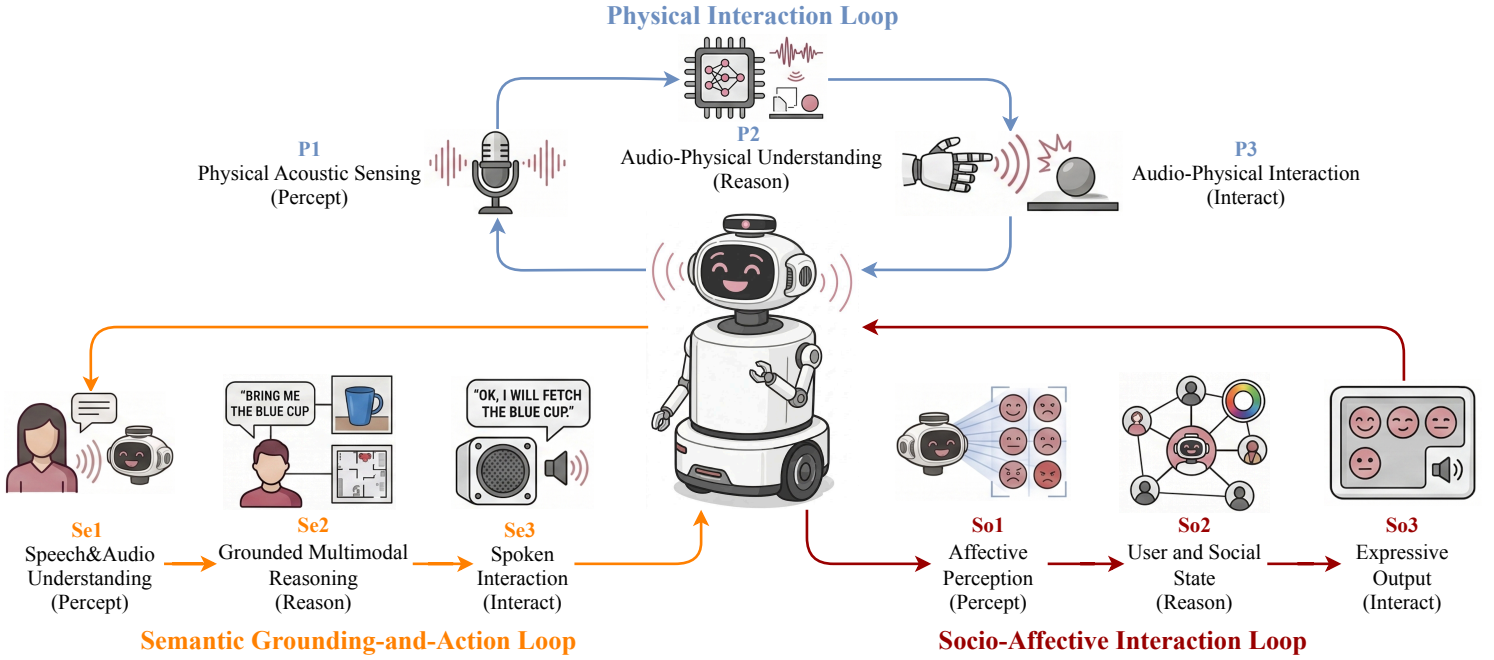


Fig. 3. The proposed unified framework for auditory-centered embodied intelligence. Three coupled loops, the Physical Interaction Loop, Semantic Grounding-and-Action Loop, and Socio-Affective Interaction Loop, span the stages of Percept, Reason, and Interact, with physical, semantic, and social information mutually informing one another.

Sections 3, 4, and 5 review the Physical Interaction Loop, the Semantic Grounding-and-Action Loop, and the Socio-Affective Interaction Loop, respectively; Section 6 summarizes datasets, benchmarks, and evaluation metrics; Section 7 discusses representative application settings; and Section 8 synthesizes open challenges and future directions.

Table 1 summarizes the resulting loop-stage categories, representative research directions, and relative maturity.

1.1 Scope and Related Surveys

Table 2 positions this work relative to five families of related reviews: robot audition and acoustic perception, embodied AI and robotics, foundation models for intelligent robots, speech and multimodal human-robot interaction, and social and affective robotics. The comparison considers whether each family explicitly addresses physical, semantic, and social roles of audio, evaluates its contribution to an embodied loop, covers foundation-model or VLA approaches, and provides an organizing taxonomy.

The comparison reveals complementary but fragmented coverage. Robot-audition surveys provide the strongest treatment of acoustic sensing, yet typically stop before grounded reasoning, action, or social interaction. General embodied-AI and foundation-model surveys cover planning, control, and language-conditioned behavior, but usually treat audition as peripheral input or preprocessing. Speech and multimodal-HRI surveys examine dialogue and interaction, while social-robotics reviews address affect, empathy, and ethics; neither tradition generally connects these concerns to physical acoustic sensing and action. Taken together, the literature lacks a common framework that follows auditory information across physical, semantic, and socio-affective roles and from perception through reasoning to interaction. This is the gap addressed by the two-axis loop-stage taxonomy proposed in this survey.

We focus primarily on work published since 2020, while including earlier studies that established foundational tasks, methods, datasets, and evaluation protocols. This emphasis reflects the recent convergence of embodied acoustic simulation, large-scale audio-visual learning,

foundation-model-based reasoning, and affect-aware human-robot interaction. The manuscript represents a June 2026 snapshot; works listed for 2026 may therefore be online preprints or early-access publications whose metadata can change before archival publication.

Our corpus deliberately excludes generic audio-language foundation models when they are not embodied, even if they are useful as upstream encoders. A general audio-text model is included only when a paper uses it in navigation, manipulation, human-robot interaction, or another robot/agent loop. This decision keeps the work focused on embodied intelligence rather than general-purpose audio representation learning.

Finally, the work is not intended to replace specialized reviews on robot audition, speech processing, audio-visual learning, or social robotics. Rather, it provides a complementary view that asks how auditory signals enter embodied loops: what is sensed, what is inferred, and how the inference changes action or interaction. Across the three loops, we compare how different technical traditions address problems within the same loop-stage taxonomy.

2 TAXONOMY AND SCOPE

2.1 Scope and Unit of Analysis

We use an *auditory embodied loop* as the unit of analysis. A work is included when sound is not merely a signal to be recognized, but part of a situated loop that connects auditory input or output to embodied perception, grounded reasoning, control, interaction, or evaluation. Under this criterion, robot audition, audio-visual navigation, contact-audio manipulation, spoken-instruction grounding, and affect-aware human-robot interaction belong to the scope because audio changes what the agent estimates, decides, says, or does.

This scope also defines the boundary of the review. Generic audio classification, speech synthesis, audio captioning, or audio-language pretraining is treated as background unless it is connected to an embodied agent, a physical environment, a robot platform, a situated task, or an action-relevant feedback process. The taxonomy therefore organizes

TABLE 1

Integrated taxonomy of multimodal auditory learning for embodied intelligence. The table summarizes the focus, research directions, and representative works associated with the nine loop–stage categories.

Category	Focus	Research directions	Representative works
Physical Interaction Loop: acoustic signals → physical state and control			
P1 Percept	Physical acoustic sensing	Sound-Source Localization and Tracking; Acoustic Event Detection; Microphone-Array Processing; Active Audition; Sound Separation and Enhancement	LOCATA (2020) [6]; ODAS (2022) [7]; Move2Hear (2021) [8]; Dynamic AV separation (2022) [9]
P2 Reason	Audio-physical understanding	Audio-Visual Navigation; Acoustic Simulation; Acoustic Mapping; Audio-Visual Scene Understanding; Sound-Based Physical-Property Estimation	SoundSpaces (2019) [10]; SoundSpaces 2.0 (2022) [11]; Waypoint AV navigation (2020) [12]; AV floorplan (2020) [13]; Sonicverse (2023) [14]
P3 Interact	Audio-physical interaction	Audio-Guided Manipulation; Audio-Driven Failure Detection; Spatial Audio Generation; Sonic Feedback Interfaces	Play it by Ear (2022) [15]; See, Hear, and Feel (2022) [16]; Audio-VLA (2025) [17]; Hearing the Robot’s Mind (2024) [18]
Semantic Grounding-and-Action Loop: Speech, sound, and language → grounded reasoning and action			
Se1 Percept	Speech and audio understanding	Automatic Speech Recognition; Spoken-Instruction Parsing; Audio Captioning; Multi-Turn Dialogue Understanding	Speech-instructed manipulation (2025) [19]; Prosodic instruction cues (2025) [20]; Language-conditioned navigation (2024) [21]
Se2 Reason	Grounded multimodal reasoning	Long-Term Memory; Audio-Language Grounding; Audio-Visual Question Answering; Audio-Conditioned VLA Reasoning	AVLMaps (2023) [22]; AVLEN (2022) [23]; NaVLA ² (2026) [24]
Se3 Interact	Spoken interaction	Full-Duplex Spoken Interaction; Refusal and Safety Explanation; Online Instruction Correction; Action-Justified Language Generation	HRI turn taking (2022) [25]; CAVEN (2024) [26]; Dialogue safety explanations (2026) [27]
Socio-Affective Interaction Loop: affective and social cues → adaptive interaction			
So1 Percept	Affective perception	Speech-Emotion Recognition; Speaker Recognition; Paralinguistic Analysis; Multimodal Affect Recognition	Personalized emotion (2024) [28]; Speech-emotion-motion (2026) [29]; Emotion-aware perception (2026) [30]
So2 Reason	User and social state	Social-Culture-Aware Reasoning; Personalized User Modeling; Relationship Modeling; Affective User-State Estimation; Privacy- and Safety-Aware Modeling	Empathy timing (2025) [31]; Trust modeling (2025) [32]; Companionship (2025) [33]; Engagement (2026) [34]
So3 Interact	Expressive output	Expressive Speech and Timing; Co-Speech Gesture Generation; Nonverbal Robot Sound; Sonified Feedback	Co-speech gesture (2025) [35]; Hearing the Robot’s Mind (2024) [18]; Consequential robot sound (2025) [36]; Nonverbal HRI sound (2023) [37]

papers by the role that audition plays in an embodied loop, rather than by audio modality, network architecture, or benchmark family alone.

Formally, this loop can be written as a three-stage mapping from situated observations to an internal state and then to action or expression:

$$\begin{aligned}
 z_t &= f_{\text{percept}}(a_{\leq t}, v_{\leq t}, l_{\leq t}, h_{\leq t}), \\
 s_t &= f_{\text{reason}}(z_{\leq t}, x_{\leq t}), \\
 (u_t, y_t) &= f_{\text{interact}}(s_t, x_t),
 \end{aligned} \tag{1}$$

where z_t denotes auditory or multimodal cues extracted from situated observations, s_t denotes task-relevant physical, semantic, or socio-affective state, u_t denotes physical action, and y_t denotes communicative or expressive output. The three functions correspond directly to the stage axis of the taxonomy: *Percept* extracts z_t , *Reason* constructs s_t , and *Interact* uses s_t to produce u_t or y_t . The loop axis specifies what s_t represents: physical state in the Physical Interaction Loop, grounded task meaning in the Semantic Grounding-and-Action Loop, and user or social state in the Socio-Affective Interaction Loop.

2.2 Stage Axis: Percept, Reason, and Interact

The proposed taxonomy is defined by two axes. The first axis specifies the stage of the embodied loop: *Percept*, *Reason*, and *Interact*. *Percept*

denotes the extraction of reliable auditory or multimodal cues from situated observations, such as source direction, event boundaries, speech content, or affective cues. *Reason* denotes the transformation of these cues into task-relevant state, grounded meaning, plans, user models, or social interpretations. *Interact* denotes the use of inferred state or meaning to produce physical action, spoken response, expressive behavior, feedback, or recovery. This distinction separates methods that share similar audio encoders but differ in the loop they close: localization as an observation, localization as evidence for navigation, and localization as a cue for recovery or user coordination are different embodied roles.

2.3 Loop Axis: Physical, Semantic, and Socio-Affective Roles

The second axis specifies the type of embodied loop in which audition is used. The *Physical Interaction Loop* addresses the role of sound as physical evidence. In this loop, auditory observations are used to infer where events occur, how objects and environments behave, and whether an ongoing action is proceeding as expected. Its progression is therefore from acoustic sensing, to audio-physical state estimation, to feedback that can alter navigation, manipulation, recovery, or human-

TABLE 2
Comparison with related survey literature across key dimensions.

Survey family	Representative surveys	Physical	Semantic	Social	Loop eval.	FMVLA	Taxonomy	Main gap
Robot audition and acoustic perception	SSL in robotics [38], [39]	✓	✗	✗	◦	✗	✗	Perception-centric; weak links to grounding, action, and social interaction.
Embodied AI and robotics	Robotics/AI reviews [1]; DMP survey [40]	◦	◦	✗	✓	◦	✗	Broad robotics scope; audition treated peripherally.
Foundation models and intelligent robots	LLM-robot surveys [2]	✗	✓	◦	◦	✓	✗	Vision/language-centric; audio remains mostly preprocessing.
Speech, dialogue, and multimodal HRI	Language/dialogue HRI [41], [42]; multimodal HRI [43], [44]	◦	✓	◦	◦	◦	✗	Interaction-centric; weak physical grounding and execution feedback.
Social robots, affect, and ethics	Empathy and ethics in social robots [45], [46]	✗	◦	✓	◦	✗	✗	Social/ethical focus; weak physical-semantic integration.
This survey	Multimodal auditory learning for embodied intelligence	✓	✓	✓	✓	✓	✓	Unified around physical, semantic, and socio-affective auditory loops.

✓ = explicitly covered; ◦ = partially covered or adjacent; ✗ = largely outside scope.

facing sonic interfaces. The detailed task families associated with this loop are summarized in Table 1.

The *Semantic Grounding-and-Action Loop* addresses the role of sound, speech, and language as carriers of task meaning. Here, the agent must recover what is being requested or described, bind that meaning to objects, places, actions, and constraints, and then produce an appropriate action or verbal response. This loop is not limited to speech recognition: it includes the broader problem of grounding auditory and linguistic evidence into situated plans, long-term context, multimodal reasoning, clarification, and action-aware language generation.

The *Socio-Affective Interaction Loop* addresses the role of auditory and expressive cues in human-centered interaction. In this loop, speech acoustics and multimodal behavior are interpreted not only as task information, but also as evidence about user state, identity, engagement, relationship, social norms, and affective context. The resulting interaction problem is to adapt the robot’s speech, gesture, facial expression, or social policy in a way that is expressive, appropriate, privacy-aware, and aligned with the ongoing relationship. Table 1 provides the corresponding task-level decomposition.

Crossing the three embodied loops with the three functional stages yields nine loop–stage categories:

- **Physical Interaction Loop** maps to **P1** *Physical Acoustic Sensing* (Percept), **P2** *Audio-Physical Understanding* (Reason), and **P3** *Audio-Physical Interaction* (Interact).
- **Semantic Grounding-and-Action Loop** maps to **Se1** *Speech and Audio Understanding* (Percept), **Se2** *Grounded Multimodal Reasoning* (Reason), and **Se3** *Spoken Interaction* (Interact).
- **Socio-Affective Interaction Loop** maps to **So1** *Affective Perception* (Percept), **So2** *User and Social State* (Reason), and **So3** *Expressive Output* (Interact).

2.4 Implications for Evaluation

These nine categories also provide a basis for evaluation. The relevant performance question is not limited to recognition accuracy; it is

whether auditory information improves closed-loop behavior, including navigation success, manipulation reliability, dialogue efficiency, safety, social appropriateness, or user experience. Evaluation should therefore include matched ablations whenever possible. For example, an audio-aware navigation policy should be compared against audio-off, vision-off, corrupted-audio, delayed-audio, or oracle-audio variants; an audio-aware manipulation policy should measure whether contact sound reduces failure latency or improves recovery; and a social robot should test whether vocal affect improves response selection without reducing user comfort or privacy.

3 PHYSICAL INTERACTION LOOP: SOUND AS PHYSICAL EVIDENCE FOR ACTION

Section 2 defines an auditory embodied loop as a progression from perceptual cues z_t , to task-relevant state s_t , and then to action or output (u_t, y_t) . The Physical Interaction Loop instantiates this formulation with sound as physical evidence. Here, z_t denotes acoustic observations such as source directions, event boundaries, separated signals, contact transients, microphone-array features, and uncertainty estimates. The inferred state s_t denotes physical beliefs such as source memory, room geometry, acoustic maps, object material, container state, contact state, or manipulation progress. The resulting action or output (u_t, y_t) may be a navigation motion, manipulation correction, failure-recovery decision, warning, spatial audio cue, or sonic feedback signal.

This loop is therefore not a collection of audio perception tasks. It describes how sound enters physical agency: first as an observation that must be stabilized under motion, ego-noise, reverberation, and overlapping sources; then as evidence for latent physical state; and finally as a signal that changes behavior or feedback. Fig. 4 should be read in this sense: it shows a progression of embodied roles rather than a fixed processing pipeline.

Physical Interaction Loop

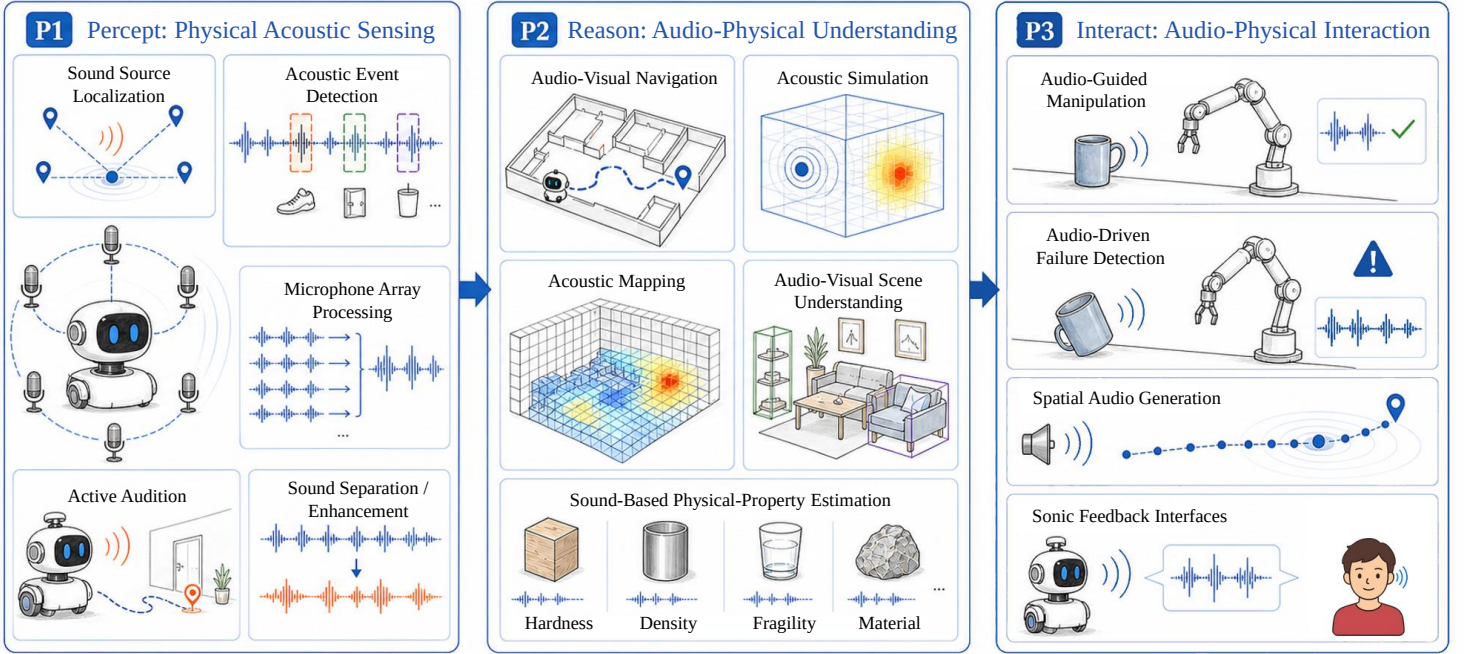


Fig. 4. Unified process view of the Physical Interaction Loop. The figure summarizes how physical acoustic sensing, audio-physical understanding, and audio-physical interaction connect auditory signals to physical state estimation and feedback.

3.1 Physical Acoustic Sensing

Physical acoustic sensing corresponds to the Percept mapping $z_t = f_{\text{percept}}(\cdot)$ in Section 2. Its goal is to convert situated sound into actionable acoustic cues, rather than to solve recognition in isolation. In the Physical Interaction Loop, these cues include source direction, event timing, separated or enhanced signals, microphone-array spatial features, contact transients, and uncertainty estimates that can support later physical reasoning or action.

Methodologically, this stage contains three recurring paradigms. Geometry-based robot audition uses microphone placement, array processing, localization, tracking, separation, and beamforming to stabilize acoustic observations. Learning-based perception estimates event boundaries, sound classes, enhancement targets, or separable sources from noisy observations. Active audition treats the robot’s motion and sensing geometry as part of the perceptual process. Representative works provide two kinds of support: LOCATA offers a benchmark for localization and tracking under dynamic source-sensor geometries, while ODAS shows how localization, tracking, separation, and post-filtering can be integrated into an embedded audition stack [6], [7].

The methodological shift is from passive acoustic estimation to active and situated sensing. Recent systems place microphones on mobile robots, UAVs, robot arms, or robot swarms, so the listening geometry itself becomes part of the problem. Outdoor HRI sound localization, humanoid array optimization, UAV sound detection, adaptive beamforming with moving objects, swarm active audition, and robotic speech enhancement with a movable microphone array all reflect this trend [47], [48], [49], [50], [51], [52]. Active audio-visual separation further shows that an embodied agent can move to improve the separability of a target sound, rather than treating separation as a fixed-sensor preprocessing step

[8], [9]. CAVER further connects sensing with exploration by letting a robot interact with objects to build audio-visual representations under uncertainty [53].

The value of physical acoustic sensing depends on whether the sensed cues can support later behavior. A source direction can guide search, a separated signal can improve speech or event recognition, an event boundary can trigger safety monitoring, and an uncertainty estimate can help the agent decide where to attend next. A recurring limitation is that many systems report localization, separation, or enhancement accuracy, but do not always measure how much these estimates improve navigation, manipulation recovery, or human-robot interaction.

3.2 Audio-Physical Understanding

Audio-physical understanding corresponds to the Reason mapping $s_t = f_{\text{reason}}(z_{\leq t}, x_{\leq t})$. The central question is how acoustic cues become beliefs about physical state. In this loop, s_t may represent a navigation belief, source memory, acoustic map, room geometry, object material, container state, contact mode, or failure hypothesis. This stage is therefore defined by interpretation: the agent must explain what physical situation produced the observed sound and how that belief should constrain future action.

The most mature line is audio-visual navigation, where an agent uses spatial audio and egocentric vision to infer room geometry, source location, and a path toward an emitting target. SoundSpaces established audio-visual navigation in acoustically simulated 3D environments, and waypoint-based navigation showed that an acoustic memory can support structured planning instead of short-horizon reactive motion [10], [12]. SoundSpaces 2.0 broadened this line by enabling continuous spatial sampling, configurable microphones and materials, and downstream

tasks such as navigation, mapping, localization, separation, acoustic matching, and far-field ASR [11].

Subsequent work expands the reasoning problem from “follow the sound” to “understand the physical situation behind the sound.” Semantic audio-visual navigation, audio-visual-language navigation, semantic-prior-based navigation, generalizable audio-visual policies, reliability-aware fusion, residual cross-modal fusion, multi-agent collaboration, and adversarial audio-visual navigation all study how audio should be fused, trusted, or ignored under occlusion, distractors, moving sources, and noisy acoustic fields [23], [54], [55], [56], [57], [58]. Sonicverse and sim-to-real acoustic field prediction emphasize a complementary question: how to build simulators and acoustic models whose reverberation, material response, and frequency-dependent fields are realistic enough for embodied learning [14], [59].

A second branch of P2 concerns physical-property and scene understanding. Here, the agent reasons about what an object or environment is like from the sound produced by interaction: material, fill state, hollowness, contact mode, room layout, or acoustically grounded scene state. Audio-visual floorplan reconstruction shows that sound can reveal geometry and room semantics beyond the camera field of view, making it a natural example of acoustic mapping and scene understanding [13]. CAVER learns audio-visual object representations through exploratory interaction, while SonicSense shows that in-hand acoustic vibration can support container-state differentiation, material prediction, shape reconstruction, and object re-identification [53], [60]. NaVLA-style multimodal instruction navigation further indicates that sound can serve as a disambiguating cue for physically grounded navigation goals [24]. The key challenge is causal ambiguity: the same sound may arise from different objects, materials, distances, or actions, so robust systems need uncertainty-aware fusion and ablations that test whether audio changes the inferred state, not only the recognition score.

3.3 Audio-Physical Interaction

Audio-physical interaction corresponds to the Interact mapping $(u_t, y_t) = f_{\text{interact}}(s_t, x_t)$. The defining criterion is that auditory evidence changes behavior or feedback. In the Physical Interaction Loop, the inferred state s_t should alter navigation, manipulation, failure recovery, spatial audio generation, or sonic feedback, rather than only improve an offline classifier. This stage is especially important for contact-rich settings because audio can reveal high-frequency events such as impact, scraping, slipping, pouring, and tool chatter before they are visually obvious.

Methodologically, P3 contains three families. Contact-audio policy learning uses sound as part of the state stream for manipulation. Audio-driven failure detection uses acoustic transients or vibration to identify slip, collision, empty pouring, or other process failures. Sonic feedback and spatial audio generation use sound as an output channel that makes robot state, event location, or task progress legible to humans or to the robot itself.

The strongest recent evidence comes from audio-guided manipulation and audio-driven failure detection. Sound-triggered mobile manipulation uses auditory events to initiate and guide task execution [61]. Play it by Ear studies audio-visual imitation learning for manipulation under occlusion, while See, Hear, and Feel shows that visual, auditory, and tactile sensing jointly improve dense packing and pouring [15], [16]. Audio-VLA injects contact audio into VLA-style robotic manipulation to perceive contact events and dynamic process feedback [17]. ManiWAV uses an ear-in-hand device to collect in-the-wild demonstrations with synchronized audio and vision, showing that contact sound helps with events, contact modes, surface materials, and

object states [62]. Hearing Touch treats contact microphones as an audio-based proxy for tactile information and transfers audio-visual pretraining to contact-rich manipulation [63]. That Sounds Right uses auditory self-supervision for dynamic manipulation, and SonicSense demonstrates that in-hand acoustic vibrations can support object perception during active exploration [60], [64]. Together, these works suggest that audio is not only an auxiliary channel; it can capture high-frequency interaction dynamics that are difficult for vision and often expensive for tactile sensing.

A complementary P3 direction studies sonic feedback and spatial audio generation for interaction. Consequential robot sounds, functional robot sounds, and explicit sonification can make contact, motion, or internal state legible to nearby users, while spatialized audio can expose where an event occurred or where attention should move [18], [36]. These interfaces are physical because they close the loop between robot action, acoustic consequence, and human or robot adjustment. Future benchmarks should therefore report action-level metrics: task success under audio removal, time-to-detect slip or collision, recovery rate after failure, unsafe-contact rate, and human understanding of robot state. A strong P3 result should show that sound changes behavior, not just that an audio encoder improves an offline classification head.

4 SEMANTIC GROUNDING-AND-ACTION LOOP: SPEECH, SOUND, AND LANGUAGE AS TASK-BEARING EVIDENCE FOR GROUNDED ACTION

Section 2 defines an auditory embodied loop as a progression from perceptual cues z_t , to task-relevant state s_t , and then to physical or communicative output (u_t, y_t) . The Semantic Grounding-and-Action Loop instantiates this formulation with speech, task-relevant sound, and language as evidence for task meaning. Here, z_t may include lexical content, candidate intents and slots, dialogue acts, prosody, stress, pauses, deictic expressions, detected sound events, temporal structure, and recognition uncertainty. The inferred state s_t may comprise grounded intent, grounded referents, spatial and temporal constraints, dialogue commitments, affordances, candidate skills, task memory, plans, and confidence or risk estimates. The output space (u_t, y_t) includes physical and communicative alternatives such as execution, clarification, confirmation, waiting, refusal, explanation, correction, and repair.

This loop is therefore not a fixed ASR-LLM-TTS pipeline. Rather, it progresses from task-bearing auditory cues through the construction of a grounded task state to the selection of physical or communicative actions. Compared with the Physical Interaction Loop, the emphasis is not only on where a sound originated or what caused it, but on what it means for the current task. Prosody is included here when it changes task interpretation or execution; affective user-state inference is treated in Section 5. Fig. 5 should be read as a progression of embodied semantic roles rather than a fixed speech-recognition-planning-response pipeline.

4.1 Speech and Audio Understanding

Speech and audio understanding corresponds to the Percept mapping $z_t = f_{\text{percept}}(\cdot)$ in Section 2. Its goal is to produce a task-bearing auditory representation rather than a transcript alone. This representation should preserve lexical, acoustic, temporal, contextual, and uncertainty information that may alter later grounding or action, while leaving the construction of a complete task state to Se2.

Methodologically, Se1 contains three recurring families. First, ASR-SLU pipelines transcribe or encode speech before estimating candidate intents, slots, commands, or semantic frames. Whisper-style ASR, SLURP, and Fluent Speech Commands provide representative foundations for robust transcription and spoken language understanding [65],

Semantic Grounding-and-Action Loop

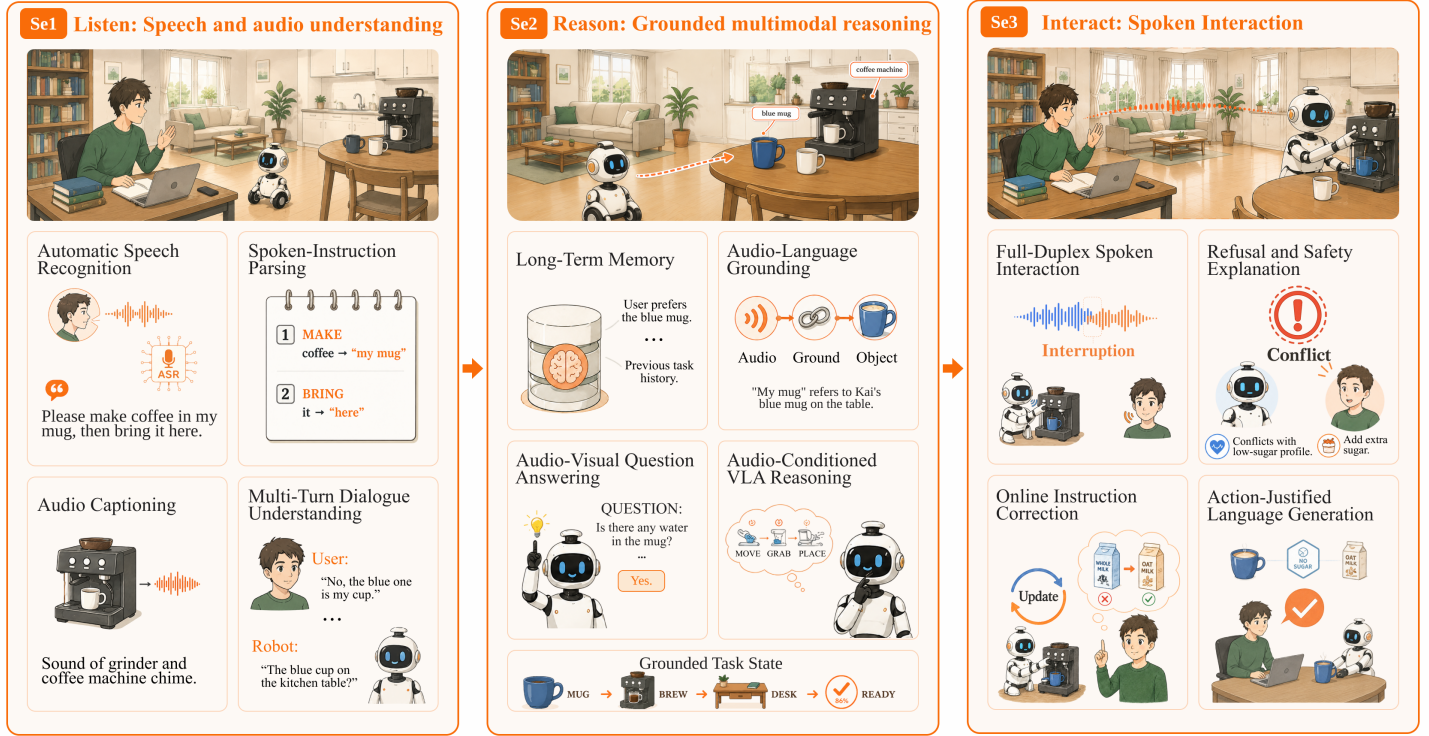


Fig. 5. Unified process view of the Semantic Grounding-and-Action Loop. The figure summarizes how speech and audio understanding, grounded multimodal reasoning, and spoken interaction connect task-bearing auditory and linguistic cues to grounded task states and physical or communicative actions.

[66], [67]. For embodied use, however, the front end must preserve action-critical entities, numbers, negation, spatial relations, temporal order, and confidence, because small errors in these elements can change a plan even when overall word error rate is low.

Second, end-to-end speech-conditioned methods retain speech representations long enough to preserve information needed by downstream task interpretation or control. VLAS integrates spoken instructions into manipulation, while speech-instructed navigation shows why spoken modifiers and preferences must be retained when they affect terrain constraints and path selection [19], [21]. Although these systems span multiple stages, their relevance to Se1 lies in preserving action-relevant speech information before grounding and control.

Third, prosody- and context-aware methods use stress, intonation, pauses, hesitation, deictic expressions, and dialogue history to preserve disambiguating evidence or produce alternative semantic hypotheses. Prosody-aware instruction understanding shows that vocal emphasis can help determine which object, relation, or action a user intends [20].

Task-relevant nonspeech events also belong to Se1 when they are detected and represented with event type, timing, and uncertainty. A doorbell, microwave beep, alarm, or ringing phone may be recognized at this stage; whether that event constitutes a navigation goal, task transition, safety constraint, or required response is determined in Se2. Estimating the physical source or cause of such a sound belongs to the Physical Interaction Loop, whereas representing it as a candidate task cue belongs here.

Se1 should be evaluated not only by transcription accuracy, but by whether action-relevant entities, constraints, prosody, event cues, and uncertainty are preserved under realistic acoustic conditions. Detailed perceptual metrics and stress tests are summarized in Section 6.

4.2 Grounded Multimodal Reasoning

Grounded multimodal reasoning corresponds to the Reason mapping

$$s_t = f_{\text{reason}}(z_{\leq t}, x_{\leq t}).$$

It maps the auditory and linguistic cues produced by Se1, together with visual observations, environment state, dialogue history, and prior actions, into a grounded task state. In this loop, s_t may contain grounded intent, grounded referents, spatial and temporal constraints, dialogue commitments, affordances, source or event memories, candidate skills, executable plans, and confidence or risk estimates.

Methodologically, Se2 contains three families. The first is structured and probabilistic grounding. Classical robotic language grounding links linguistic constituents to objects, places, paths, events, and actions under uncertainty [41], [68], [69]. R2R and ALFRED later operationalized grounded instruction following through partial observability and sequential action [70], [71]. Although these benchmarks are not primarily auditory, they define embodied grounding settings into which speech and task-relevant sound must be integrated.

The second family is spatial, event-memory, and assistance-aware grounding. This family is especially important for auditory embodied intelligence because sounds are often intermittent, outside the field of view, or separated in time from the action they should influence. AVLMaps binds audio, vision, and language in a spatial representation so that sound snippets and language queries can refer to sounding objects, events, and locations [22]. AVLEN combines audio-visual navigation with language assistance and query decisions, making auditory evidence relevant both to spatial belief and help seeking [23]. CAVEN extends conversational navigation to noisy environments in which an agent must

decide how to approach an audio goal and when dialogue is needed [26]. NaVLA² moves toward a unified vision-language-audio-action formulation for multimodal instruction navigation [24]. Together, these works motivate grounded states that retain time-stamped event memory, object–place bindings, source uncertainty, task phase, dialogue history, and the provenance of auditory evidence.

The boundary between the Physical and Semantic Interaction Loops remains important. A physical module may infer that a click indicates successful insertion, that an alarm originates in another room, or that an impact was caused by a dropped object. Semantic reasoning then determines whether the insertion subtask is complete, the alarm introduces a safety constraint, or the dropped object should become the next recovery goal. Sound becomes semantic evidence when a physically interpreted event revises task state, memory, plan, or action priority.

The third family is foundation-model and VLA reasoning. SayCan constrains language-model skill proposals with robot affordances, while Inner Monologue uses environmental and human feedback for iterative replanning [72], [73]. Representative multimodal and VLA systems such as RT-2, PaLM-E, and OpenVLA further shift the field toward generalist skill selection, embodied representations, and language-conditioned policy learning [3], [4], [5]. These systems are not primarily auditory, but they provide architectures in which speech and sound can enter as state evidence. Their relevance to this survey lies in whether auditory information becomes a first-class part of policy-level reasoning rather than remaining ASR preprocessing.

The central Se2 criterion is causal contribution: removing, corrupting, or delaying a task-relevant auditory cue should measurably alter grounding, confidence, memory, or planning, whereas irrelevant audio should not dominate the decision. A strong system should also revise an incorrect auditory hypothesis when later observations or feedback contradict it.

4.3 Spoken Interaction

Spoken interaction corresponds to the Interact mapping

$$(u_t, y_t) = f_{\text{interact}}(s_t, x_t).$$

It closes the semantic loop by selecting among physical and communicative actions. Given a grounded task state, an agent may execute a plan, wait, ask for clarification, confirm a referent, refuse an unsafe or impossible request, explain uncertainty, accept a correction, repair a failed action, or request human assistance. These are embodied policy choices under uncertainty and risk, not peripheral language-generation functions.

Methodologically, Se3 contains three families. The first is streaming and turn-taking coordination. Endpointing, turn-taking, overlap handling, interruption detection, barge-in recovery, and response latency determine whether a robot can coordinate speech with ongoing action. HRI turn-taking work shows how multimodal conversational cues can support decisions about when to take, hold, or yield the floor [25]. Full-duplex interaction further requires the agent to coordinate listening, speaking, interruption handling, and physical action in real time.

The second family is clarification, correction, and help seeking. TEACH connects dialogue, visual state, and household task execution; DialFRED makes clarification part of embodied instruction following; and TEACH-DA links dialogue acts such as request, confirm, acknowledge, and correct to task progress [74], [75], [76]. These benchmarks are not primarily auditory, but they define embodied interaction settings into which spoken language must be inserted. AVLEN and CAVEN provide more explicitly auditory examples in which an agent balances navigation

progress against the cost of asking for help [23], [26]. The central act-versus-ask decision is whether clarification is expected to prevent a costly error without introducing unnecessary queries when the grounded state is already reliable.

The third family is refusal, explanation, and action repair. Unsafe, infeasible, contradictory, or weakly grounded requests may require refusal or escalation rather than execution. After failure, the agent should report what happened, expose relevant uncertainty, update common ground, and replan. Dialogue-based safety explanations and evidence-grounded HRI with persistent memory illustrate how explanations can be tied to observations and action history rather than generated as detached text [27], [77].

Se3 should be judged by whether interaction prevents unsafe or incorrect actions, enables timely clarification and repair, and improves task progress with acceptable latency and user burden. Detailed interaction metrics are summarized in Section 6.

Overall, the Semantic Grounding-and-Action Loop converts task-bearing speech, sound, and language into grounded state and then into a selection over physical and communicative actions. Its main open gap is that auditory evidence still too rarely changes the agent’s belief, policy, and recovery mechanism in a causal and transparent way.

5 SOCIO-AFFECTIVE INTERACTION LOOP: AFFECTIVE AND SOCIAL CUES AS EVIDENCE FOR ADAPTIVE INTERACTION

Section 2 defines an auditory embodied loop as a progression from perceptual cues z_t , to task-relevant state s_t , and then to physical or communicative output (u_t, y_t) . The Socio-Affective Interaction Loop instantiates this formulation with affective, paralinguistic, and interactional cues as evidence for human-centered adaptation. Here, z_t may include prosody, hesitation, laughter, silence, overlap, turn timing, and multimodal synchrony. The inferred state s_t is a bounded user and social belief state that may include engagement, confusion, trust, user preference, relationship history, applicable norms, and uncertainty. The output (u_t, y_t) may include physical choices such as pausing, assisting, yielding, or waiting, together with expressive choices such as speech style, gaze, gesture, facial expression, nonverbal sound, or sonified feedback.

This loop is not a fixed emotion-recognition–user-state-classification–emotion-generation pipeline. It progresses from uncertain affective evidence, through a temporally grounded and norm-constrained social state, to adaptive behavior. Section 4 treats speech and sound primarily as task-bearing evidence; this section focuses on how the user participates in interaction and how the robot should adapt its timing, style, channel, and intensity. Prosody belongs to Section 4 when it changes task meaning or grounding, and to this loop when it indicates engagement, urgency, confidence, or comfort. Fig. 6 should be read as a progression of socio-affective embodied roles rather than a fixed processing pipeline.

5.1 Affective Perception

Affective perception corresponds to the Percept mapping $z_t = f_{\text{percept}}(\cdot)$ in Section 2. Its goal is to extract reliable affective, paralinguistic, and interactional cues from situated multimodal observations. The output of So1 should not be a definitive assertion that a user is happy, angry, anxious, or disengaged. Instead, it should preserve uncertain evidence for competing interpretations in So2, because a pause, raised voice, laugh, or interruption depends on the speaker, task phase, acoustic environment, and recent interaction.

Methodologically, So1 contains three recurring families. The first is acoustic and paralinguistic affect sensing. These methods estimate

Socio-Affective Interaction Loop

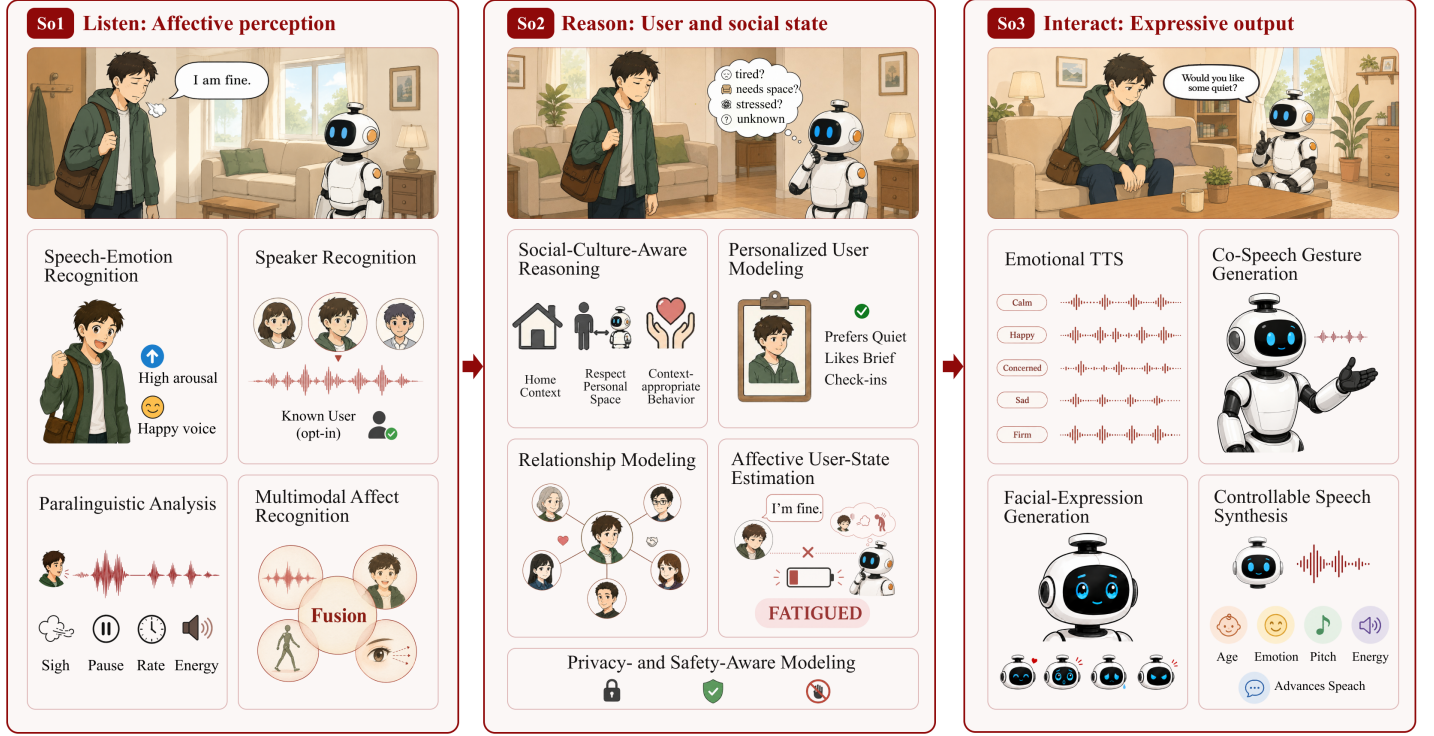


Fig. 6. Unified process view of the Socio-Affective Interaction Loop. The figure summarizes how affective perception, user and social state reasoning, and expressive output connect socio-affective cues to a bounded social belief state and adaptive interaction.

discrete affect categories, continuous dimensions such as arousal and valence, or observable interactional cues from prosody, voice quality, speaking rate, hesitation, laughter, silence, overlap, and turn timing. Speaker attribution is relevant when it supports speaker-conditioned calibration, helping distinguish personal vocal baselines from interaction-related changes rather than enabling unrestricted identity inference.

The second family is multimodal affect and engagement perception. Auditory evidence is fused with facial expression, gaze, gesture, posture, touch, dialogue context, and temporal synchrony because individual cues are often ambiguous. Speech-emotion-motion and emotion-aware gesture systems illustrate how vocal, visual, and motion representations can be connected in embodied platforms [29], [30]. Although these systems span perception and expression, their relevance to So1 lies in extracting affective evidence before a response is selected.

The third family is personalized and uncertainty-aware perception. Vocal expression varies across speakers, languages, accents, cultures, devices, and environments, so a universal acoustic-to-social mapping is rarely reliable. Personalized speech emotion recognition in HRI demonstrates the value of adapting to individual vocal characteristics [28]. A practical system should expose calibrated confidence, represent disagreement across modalities, and abstain when evidence is weak or contradictory.

The embodied setting adds far-field microphones, reverberation, ego-noise, overlapping speakers, moving users, and the robot's own sounds. A strong So1 system should preserve calibrated evidence under these shifts and show that the cues improve later timing, comfort, clarification, or social appropriateness. Evaluation should therefore include speaker and domain shift, overlap, confidence calibration, privacy-preserving processing, and downstream interaction effects; detailed perceptual

metrics remain in Section 6.

5.2 User and Social State

User and social state corresponds to the Reason mapping $s_t = f_{\text{reason}}(z_{\leq t}, x_{\leq t})$ in Section 2. It transforms affective cues, dialogue history, task progress, environmental context, prior interaction, and user preferences into an action-relevant social belief state. This state should distinguish transient conditions, such as attention, confusion, stress, or willingness to continue, from stable user preferences and relationship-level variables such as trust, rapport, social role, and interaction history.

Methodologically, So2 contains three recurring families. The first is temporal user-state estimation. A pause may indicate reflection, uncertainty, politeness, distraction, or sensing failure, while a raised voice may reflect urgency, excitement, frustration, or noise. Social reasoning must therefore integrate evidence across turns, represent alternative hypotheses, and revise them when later observations conflict with an earlier interpretation. The goal is not a fixed emotion label, but a temporally consistent belief about interaction conditions relevant to action.

The second family is personalized and relational modeling. A user model may represent preferred explanation detail, interaction pace, accessibility needs, or tolerance for proactive assistance. A relationship model may retain prior corrections, failures, successful adaptations, and changes in trust or rapport across sessions. Work on enculturated companion design emphasizes that personalization and perceived care must be situated within the user's social and cultural context [32]. Longitudinal studies also show that immediate engagement may not transfer across settings or persist after the robot is removed, making long-term evaluation necessary [34].

The third family is norm-, culture-, privacy-, and safety-constrained reasoning. A robot should not adapt merely because an affect classifier produces a confident score. It must consider whether the inference is relevant, whether the context is sensitive, whether the user has consented to personalization, and whether asking directly would be safer. Social roles and cultural conventions affect silence, gaze, interruption, emotional intensity, and empathy. Work on when and how robots should express empathy shows that empathy is a timing and policy problem: clarification or a neutral response may better preserve user agency [31].

A useful So2 representation is therefore a bounded belief state over interaction hypotheses, including short-term state, longer-term preferences, relationship history, applicable norms, confidence, and evidence provenance. It should also support non-inference: when evidence is weak, culturally ambiguous, privacy-sensitive, or irrelevant, the agent may maintain its current policy, ask directly, or avoid personalization.

A strong So2 system should maintain temporally consistent and calibrated beliefs, improve adaptation relative to history-free or non-personalized baselines, and recover after an incorrect hypothesis. Evaluation should include personalization gain, memory consistency, uncertainty calibration, cultural and language variation, consent and privacy cost, and longitudinal effects on trust, comfort, engagement, and task progress. It should also test whether later evidence corrects an initially wrong social hypothesis. Adaptation-off and memory-off ablations are important for establishing causal contribution.

5.3 Expressive Output

Expressive output corresponds to the Interact mapping $(u_t, y_t) = f_{\text{interact}}(s_t, x_t)$ in Section 2. It closes the socio-affective loop by using the inferred state to modulate physical and communicative behavior. So3 concerns response timing, channel, style, and intensity. The central question is not whether output appears emotional in isolation, but whether it is warranted, understandable, appropriately timed, safe, and aligned with the robot’s capabilities.

Methodologically, So3 contains three recurring families. The first is expressive speech and interaction timing. A robot can adapt speaking rate, intensity, utterance length, pauses, backchannels, endpointing, and latency. It may slow down when the user appears confused or issue a concise warning in a safety-critical situation. When the state estimate is uncertain, expressive speech should invite correction rather than assert an inferred emotion as fact.

The second family is coordinated multimodal expression. Speech style can be synchronized with facial expression, gaze, head movement, gesture, posture, and, where appropriate and consented to, affective touch. Co-speech gesture generation connects semantic and affective information with humanoid motion [35]. Work on touch, robot-led emotion regulation, and adaptive facial expression further illustrates the shift from static displays toward interaction-dependent behavior [78], [79], [80]. The channels should jointly communicate attention, uncertainty, encouragement, warning, or turn ownership without exaggerating the robot’s internal state.

The third family is nonverbal sound and sonified feedback. Designed tones, earcons, spatialized sound, movement sound, and sonification can communicate waiting, uncertainty, completion, warning, or navigation intent when visual attention is divided. Work on explicit sonification, consequential and functional robot sounds, and nonverbal sound in HRI shows that auditory output can improve legibility and coordination, but can also create annoyance, startle, or misleading interpretations [18], [36], [37].

The boundary with the other loops is important. Se3 determines the task-level communicative act—whether to execute, ask, confirm,

refuse, explain, or repair—whereas So3 determines how that act is timed and expressed. Likewise, sound that communicates contact failure or physical state belongs primarily to P3, while sound regulating attention, turn-taking, comfort, trust, or social coordination belongs to So3.

Expressive output is therefore a bounded social control problem, not merely emotional imitation. A robot should not maximize apparent empathy or intensity; a neutral or less expressive response may be more appropriate. Strong So3 systems should be compared with style-neutral, expression-off, and sonification-off baselines and evaluated through timing, coordination, cross-modal consistency, comfort, annoyance, appropriateness, trust calibration, uncertainty communication, and perceived safety.

Overall, the Socio-Affective Interaction Loop reframes social audio from isolated emotion recognition to uncertainty-aware interaction regulation. The agent extracts affective evidence, constructs a temporally grounded and norm-constrained social belief state, and uses that state to modulate physical and expressive behavior. The desired outcome is calibrated, transparent, privacy-aware, and socially bounded interaction.

6 DATASETS, BENCHMARKS, AND METRICS

This section summarizes the datasets, benchmarks, and metrics for auditory embodied intelligence. These three elements play different roles: datasets provide sensory observations and annotations, benchmarks define tasks and protocols, and metrics determine what counts as progress. Due to the scarcity of datasets covering complete closed-loops, we organize the discussion around the three stages. At the *Percept* stage, evaluation asks whether an agent can extract reliable auditory or multimodal cues. At the *Reason* stage, it asks whether these cues improve state estimation, grounding, planning, or user-state inference. At the *Interact* stage, it asks whether auditory information meaningfully changes action, dialogue, feedback, safety, or social behavior. Table 3 summarizes representative datasets and benchmarks, and Table 4 summarizes metrics already reported in the literature as well as those still needed for closed-loop auditory embodied agents.

6.1 Datasets and Benchmarks

Datasets provide the sensory observations and annotations from which auditory embodied agents learn, while benchmarks turn those resources into comparable tasks, splits, protocols, and success conditions. We discuss them together because many resources serve both roles: they define what is recorded and also how progress is measured. A useful resource should record not only what the agent hears, but also the situation in which the sound occurs, including microphone geometry, robot or human actions, visual context, task or dialogue state, and the interaction partner. Across the three stages, the central difference is whether the resource labels perceived signals, inferred states, or behavioral consequences.

Percept-stage resources. Percept-stage datasets and benchmarks evaluate the listening front end before the agent reasons or acts. They ask whether auditory or multimodal cues can be extracted reliably under noise, reverberation, device variation, motion, and domain shift. *Acoustic Perception* resources focus on event classes, acoustic scenes, source locations, and temporal activity. AudioSet [81] represents large-scale sound-event supervision, while VGGSound [87] connects visible objects and actions with sound. Spatial and embodied listening resources such as LOCATA [6], STARSS23 [90], and Real Acoustic Fields [111] stress localization, room transfer, distributed sensing, and real acoustic fields; BatVision [116] further illustrates active echo-based egocentric perception. Table 3 gives the fuller list of event, scene, room-acoustic, robot-listening, and transfer benchmarks. *Semantic Perception*

TABLE 3

Summary of datasets, benchmarks, and resources in Section 6. Rows follow the three-stage taxonomy; each work is listed once under its primary role.

Stage	Category	Datasets / benchmarks / resources	Main use
Percept: extract reliable auditory or multimodal cues			
Percept	Acoustic perception	AudioSet [81]; FSD50K [82]; ESC [83]; UrbanSound [84]; TUT/RealISED [85], [86]; VGGSound/VGG-SS [87], [88]; LOCATA [6]; TAU-NIGENS/DCASE SELD [89]; STARSS23/DCASE SELD [90]; DREGON [91]; UaVirBASE [92]; Pepper SSLR/AV16.3/NAL-NAR [93], [94], [95]; CHiME-6/DiPCo [96], [97]; DIRHA/VoiceHome [98], [99]; RealMAN/Massive distributed microphone array [100], [101]; dEchorate/OpenAIR/Aachen/MeshRIR/BUT Reverb/ACE/MIRACLE/SRIRACHA/MP-RIR/RAF [102], [103], [104], [105], [106], [107], [108], [109], [110], [111]; CIPIC/VCL [112], [113]; Urbansas/StreetAware [114], [115]; BatVision [116]; HEAR [117]	Sound events, scenes, audio-visual events, localization/tracking, UAV and robot listening, meeting/home arrays, ego-noise, RIR/HRTF and room transfer, active echolocation, bioacoustics, urban multimodal tracking, and representation transfer.
Percept	Semantic perception	Speech Commands [118]; SLURP [66]; LibriSpeech [119]; AISHELL-1/2 [120], [121]; Common Voice [122]; TED-LIUM [123]; Wall Street Journal [124]; THCHS-30 [125]; AudioCaps [126]; Clotho [127]; WavCaps [128]; MACS [129]; GTZAN [130]; SUPERB [131]; AudioBench [132]; AIR-Bench [133]	Commands, intents/slots, ASR, captions, music semantics, speech representation, and audio-language understanding.
Percept	Physical perception	The Greatest Hits [134]	Observable contact and material cues from impact/scraping sounds in video.
Percept	Social perception	IEMOCAP [135]; RAVDESS [136]; CREMA-D [137]; Emo-DB [138]; VoxCeleb1/2 [139], [140]; Speakers in the Wild [141]; VCTK [142]; L2-ARCTIC [143]; Common Voice/AISHELL-style corpora [120], [121], [122]; AVEC 2019 [144]	Emotional speech, speaker identity, accent/language/channel metadata, paralinguistic affect, and front-end affect sensing.
Reason: use sound as evidence for state, answer, plan, or explanation			
Reason	Acoustic reasoning	SoundSpaces [10]; SoundSpaces 2.0 [11]; Sonicverse [14]; BeDAViN [145]; Hear You Are QA [146]; AVLEN [23]; AVLMaps [22]; FUSS [147]; Learning to Set Waypoints [12]; Semantic Audio-Visual Navigation [148]; Audio-Guided Visual Perception [149]; audio-visual floorplan reconstruction [13]; NaVLA-style navigation [24]	Audio-goal navigation, spatial belief, acoustic memory, noisy multi-source navigation, relative spatial QA, mixture/source reasoning, map-based audio-language grounding, intermittent sound goals, and language-assisted navigation.
Reason	Semantic reasoning	MUSIC-AVQA [150]; Clotho-AQA [151]; AVSD [152]; MMAU [153]; PAQA [154]; DCASE 2025 Bioacoustic/Temporal Soundscape AQA [155]; CompA [156]; GAMA/CompA-R-test [157]; MMAR [158]; OmniBench [159]; MMSU [160]; MSU-Bench [161]; Audio Reasoning Challenge [162]	Audio QA, multi-round and multi-domain audio QA, bioacoustic and temporal soundscape QA, audio-visual dialogue, compositional audio-language reasoning, spoken-language reasoning, and rationale evaluation.
Reason	Physical reasoning	ObjectFolder [163]; ObjectFolder 2.0/Real [164], [165]; RealImpact [166]; RSAudio [167]; CAVER [53]; Fallen Objects [168]; PhyAVBench [169]	Object/material state, impact sound fields, multisensory object reasoning, exploratory object learning, asynchronous out-of-view impact inference, and audio-physics sensitivity.
Reason	Social reasoning	CMU-MOSI [170]; CMU-MOSEI [171]; MELD [172]; M3ED [173]; MUSTARD [174]; MIntRec [175]	Contextual sentiment, emotion, sarcasm, and multimodal intent inference.
Interact: evaluate behavior after listening and reasoning			
Interact	Acoustic interaction	Move2Hear [8]; Active Audio-Visual Separation [9]	Active movement for better localization/separation under changing mixtures.
Interact	Semantic interaction	AVN-Instruct/CAVEN [26]; HRI Turn-Taking Corpus [25]; Fluent Speech Commands [176]; Snips SmartLights/SmartSpeaker (EN/FR) [177]	Conversational navigation, query decisions, spoken turn timing, repair, keyword and intent/slot grounding, and smart-home or smart-speaker command grounding.
Interact	Physical interaction	Play it by Ear [15]; That Sounds Right [64]; See, Hear, and Feel [16]; Hearing Touch [63]; ManiWAV [62]; Audio-VLA [17]; SonicSense [60]; AudioPouring/PouringNet [178]; CMU Food Manipulation [179]; Multi-Keyboard Acoustic [180]; emotional knocking-sound dataset [181]	Contact-audio manipulation, failure detection, object state, vibration sensing, liquid/granular and deformable-object state, policy improvement, operation sounds, and contact-action cues.
Interact	Social and safety interaction	SEWA DB [182]; K-EmoCon [183]; NoXi+J [184]; GroupAffect-4 [185]; Xperience-10M [186]; CASIA Chinese Emotional Speech [187]; EMOVIE [188]; Hearing the Robot's Mind [18]; SafeVLA-Bench [189]; Beyond Task Success [190]	Social response, engagement, egocentric interaction streams, Mandarin emotional speech/dialogue, explicit feedback, unsafe actions, and behavioral diagnostics.

resources connect audio to symbolic content. Speech Commands [118] and SLURP [66] cover command and intent understanding, LibriSpeech [119] anchors ASR-style supervision, and AudioCaps [126] and Clotho [127] map environmental audio to captions. Transfer benchmarks such as SUPERB [131], AudioBench [132], and AIR-Bench [133] test whether such representations generalize across speech, paralinguistic, environmental, and music tasks. *Physical Perception* resources capture directly observable physical cues in sound. The Greatest Hits [134] studies sounds produced by striking and scraping objects in video, making it useful for learning contact and material cues before moving to state reasoning or control. *Social Perception* resources represent how people sound and behave in affective displays, speaker recordings, or dyadic interaction. IEMOCAP [135] and RAVDESS [136] are typical emotional-speech resources, while VoxCeleb1/2 [139], [140] support speaker recognition. AVEC 2019 [144] illustrates how such front-end cues can be benchmarked for state-of-mind recognition, depression assessment, and cross-cultural affect sensing.

Reason-stage resources. Reason-stage resources go beyond recognizing what is heard. They ask whether audio changes an inferred state, answer, plan, or physical explanation under ambiguity. *Acoustic Reasoning* resources are led by audio-visual navigation and spatial inference. SoundSpaces [10], SoundSpaces 2.0 [11], Sonicverse [14], and BeDAViN [145] place agents in simulated or benchmarked 3D environments where audio carries information about source location, room geometry, reverberation, distractors, and navigation goals. Hear You Are QA [146] turns simulated visual-acoustic scenes into spatial question answering, while AVLEN [23] adds language feedback and query decisions. The fuller set in Table 3 shows how this line expands from source localization to acoustic memory, map-based grounding, and help-seeking behavior. *Semantic Reasoning* resources use audio as evidence for questions, dialogue, and multi-step inference. MUSIC-AVQA [150], Clotho-AQA [151], and AVSD [152] cover early audio-visual QA and dialogue settings, while newer benchmarks such as MMAU [153], CompA [156], and MMAR [158] make compositional, open-ended, and

TABLE 4
Metrics for the three-stage auditory embodied evaluation pipeline. The table groups the metrics used by the cited works.

Stage	Category	Metric sources / protocols	Core metrics	Embodied diagnostics
Percept: quality of auditory observations				
Percept	Acoustic perception	LOCATA [6]; STARSS23/DCASE SELD [90]; PSDS [191]; BSS Eval [192]; SI-SDR/STOI/PESQ/DNSMOS/ViSQOL [193], [194], [195], [196], [197]	DOA/localization error; tracking continuity; accuracy/F1/mAP; event or segment ER; PSDS; SELD error/F1/LE/LR; SDR/SIR/SAR; SI-SDR; intelligibility and perceptual quality.	Latency, real-time factor, microphone geometry, ego-noise, robot motion, reverberation, source overlap, threshold sensitivity, and device shift.
Percept	Semantic perception	Speech Commands [118]; SLURP [66]; ASR semantic/CER studies [198], [199], [200]; Clotho/AudioCaps/WavCaps [126], [127], [128]; MACE [201]; SUPERB/HEAR/AudioBench/AIR-Bench [117], [131], [132], [133]	WER/CER; entity or semantic WER; command accuracy; intent accuracy; slot F1; BLEU/METEOR/CIDEr/SPICE/SPIDER; MACE; task-family score.	Noise/accent/prosody shift, streaming ASR stability, instruction ambiguity, entity preservation, caption grounding, and audio-language calibration.
Percept	Physical perception	The Greatest Hits [134]	Contact/event class; temporal alignment; material or action recognition.	Generalization across objects, surfaces, tools, and unseen contact conditions.
Percept	Social perception	IEMOCAP [135]; RAVDESS [136]; AVEC 2019 [144]	Accuracy/UAR; macro-F1; MAE/RMSE; correlation; CCC for arousal/valence or health state; speaker verification or diarization error where relevant.	Speaker/accent/culture/language slices, overlap speech, privacy, fairness gap, and confidence calibration under domain shift.
Reason: effect of audio on state, answer, plan, or explanation				
Reason	Acoustic reasoning	SoundSpaces family [10], [11], [12]; Semantic Audio-Visual Navigation [148]; AVLEN [23]; embodied navigation evaluation [202]	SR; SPL; path length; DTG; collision rate; SNA; SWS; query rate; unseen-scene/sound generalization.	Audio-off, corrupted/delayed audio, distractor sources, moving goals, oracle audio, and real-room transfer.
Reason	Semantic reasoning	MUSIC-AVQA [150]; Clotho-AQA [151]; AVSD [152]; CompA/GAMA [156], [157]; MMAU [153]; MMAR [158]; OmniBench [159]; MMSU/MSU-Bench [160], [161]; Audio Reasoning Challenge [162]	EM/token F1; answer or MCQ accuracy; temporal/counting accuracy; subcategory and compositional scores; dialogue score; rationale factuality and logic.	Text-only, audio-shuffle, counterfactual-audio, and hard-negative tests; abstention accuracy, hallucination rate, confidence calibration, temporal evidence, and answer-rationale consistency.
Reason	Physical reasoning	ObjectFolder family [163], [164]; PhyAVBench [169]	Object/material accuracy; contact-state error; geometry/source-state error; audio-physics sensitivity; contrastive physical response.	Counterfactual physical prompts, object shift, sim-to-real gap, and causal audio contribution.
Reason	Social reasoning	CMU-MOSI/MOSEI [170], [171]; MELD/M3ED [172], [173]; MUsTARD [174]; MIntRec [175]	Sentiment/emotion/intent accuracy or F1; arousal-valence error; turn-level consistency; sarcasm or intent score.	History consistency, personalization gain, uncertainty, fairness, and robustness to conversational context shift.
Interact: behavioral consequence of listening				
Interact	Acoustic interaction	Move2Hear [8]; Active Audio-Visual Separation [9]	Separation gain; localization gain; target-to-interferer improvement; time-to-acquire; movement cost.	Moving-source stress tests, action efficiency, collision/safety during listening, audio removal or delay, and compute/energy cost.
Interact	Physical interaction	Play it by Ear [15]; That Sounds Right [64]; See, Hear, and Feel [16]; Hearing Touch [63]; ManiWAV [62]; Audio-VLA [17]; SonicSense [60]	Task success; TCR; contact-state accuracy; failure-detection time; recovery rate; force/pose error; unsafe-contact rate.	Audio ablation, delayed contact cues, recovery after slip/collision, policy latency, and trajectory change caused by sound.
Interact	Semantic interaction	AVN-Instruct/CAVEN [26]; HRI turn taking [25]; PARADISE/dialogue evaluation [203], [204]	Clarification/correction success; act/ask/refuse accuracy; turn-taking F1; endpointing and first-response latency; barge-in recovery; coherence; explanation faithfulness.	Cost of asking, human response latency, over-querying, repair turns, early/late cutoff, spoken-command grounding under noise, and user satisfaction.
Interact	Social and safety interaction	SEWA DB [182]; K-EmoCon [183]; NoXi+J [184]; GroupAffect-4 [185]; Godspeed [205]; spoofing/security metrics [206]; Hearing the Robot's Mind [18]; SafeVLA-Bench [189]; Beyond Task Success [190]	Trust; rapport; comfort; engagement; perceived intelligence/safety; satisfaction; SBU; VSI; spoofing or attack success; rollout failure signatures; runtime cost.	Consent, long-term adaptation, manipulation risk, cultural variation, privacy leakage, uncertainty communication, and human-rated safety.

rationale-based reasoning more explicit. These resources shift evaluation from recognizing acoustic content to justifying answers over temporal, semantic, and multi-source evidence. *Physical Reasoning* resources connect sound to latent object properties and cross-modal object state. ObjectFolder [163], ObjectFolder 2.0 [164], and ObjectFolder Real [165] provide multisensory object representations, real object impacts, and tactile-video observations, while RealImpact [166] records impact sound fields with physical and visual metadata. RSAudio [167] adds synthetic audio-visual impact examples for shape and material inference. Together these resources ask whether sound can reveal material, geometry, contact, or out-of-view physical state rather than only event identity. *Social Reasoning* resources use speech audio together with language, face, body, or context to infer human state. CMU-MOSI [170], CMU-MOSEI [171], and MELD [172] illustrate the move from clip-level affect toward contextual sentiment, emotion, and dialogue understanding. These resources show that social sound is most useful when interpreted with context and interaction history.

Interact-stage resources. Interact-stage resources evaluate behavior

after listening and reasoning. They should include audio together with actions, feedback, recovery, dialogue turns, or user responses; because public closed-loop auditory benchmarks are still rare, current evidence often comes from task protocols in robot manipulation, conversational navigation, active listening, and HRI feedback studies. *Acoustic Interaction* resources test whether movement improves what the agent can hear. Move2Hear [8] and Active Audio-Visual Separation [9] evaluate active listening under changing mixtures and moving sources. *Semantic Interaction* resources connect audio with task dialogue and decisions about when to act or ask. AVN-Instruct/CAVEN [26] evaluates navigation toward audio goals with help-seeking decisions, and the HRI Turn-Taking Corpus [25] provides spoken human-robot dialogue cues for deciding when to take, hold, or yield the floor. Command corpora remain useful as weakly interactive grounding resources; representative examples are listed in Table 3. *Physical Interaction* resources test whether contact sound improves action. Play it by Ear [15] studies audio-visual imitation under occlusion, while That Sounds Right [64], ManiWAV [62], SonicSense [60], and Audio-VLA [17] evaluate whether contact audio or vibration improves object-state inference, process monitoring, and ma-

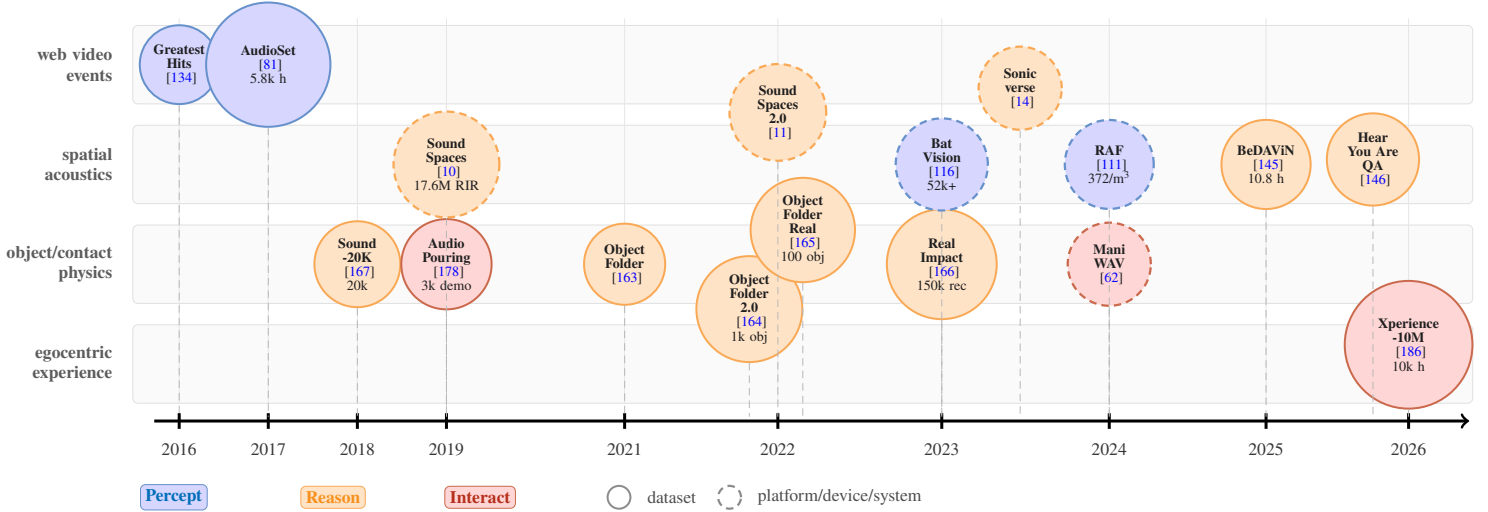


Fig. 7. Embodied audio dataset timeline. Colors indicate the primary stage. Circle size encodes the closest reported scale proxy: hours/clips for web or egocentric data, room impulse response density (RIR) density for acoustics, number of recordings, objects, demonstrations, or episodes for embodied resources. Vertical bands indicate data context, and dashed borders mark platforms, devices, or system-style resources.

nipulation success. AudioPouring-style resources [178] show the same idea in liquid-level and flow estimation. The important distinction from physical reasoning is causal: audio should change the action trajectory, not only improve an offline classifier. *Social and Safety Interaction* resources capture how people respond to one another and how agents communicate uncertainty, safety, or intent. SEWA DB [182] and K-EmoCon [183] provide multimodal affect and engagement signals, while Xperience-10M [186] represents a broader egocentric interaction stream with synchronized video, depth, pose, motion, IMU, audio, and language annotations. HRI and VLA-adjacent benchmarks such as SafeVLA-Bench [189] and Beyond Task Success [190] identify diagnostics for unsafe actions, confidence, latency, and behavioral failure signatures. A mature auditory Interact-stage benchmark should evaluate not only final success, but also timing, recovery, uncertainty communication, safety, and the effect of removing or corrupting audio.

6.2 Metrics

Metrics define what counts as progress. For auditory embodied intelligence, they should answer three progressively stronger questions. First, at the Percept stage, did the agent extract the right auditory observation under realistic acoustics? Second, at the Reason stage, did this observation change the inferred state, answer, plan, or explanation in a justified way? Third, at the Interact stage, did the use of sound improve action, dialogue, recovery, safety, or human experience? Table 4 therefore groups metrics by stage and by the role of audio. The table also separates *core task metrics* from *embodied diagnostics*, because high offline accuracy can hide failures caused by ego-noise, latency, user interruption, domain shift, or unsafe action.

Percept-stage metrics. Percept-stage metrics evaluate the listening front end before its output is used for planning or interaction. *Acoustic Perception* should combine spatial, event, separation, enhancement, and runtime measures. Direction-of-arrival and localization error, tracking continuity, localization recall, and SELD metrics evaluate where and when sounds occur in protocols such as LOCATA and DCASE/STARSS23 [6], [90]. Sound-event and scene analysis should report accuracy, macro/micro F1, mean average precision, event- or segment-based error rate, and threshold-robust scores such as PSDS [191]. Separation and enhancement should not rely on a single signal score: SDR/SIR/SAR from BSS

Eval [192], SI-SDR [193], intelligibility metrics such as STOI [194], and perceptual quality metrics such as PESQ, DNSMOS, or ViSQOL [195], [196], [197] capture complementary failure modes. For embodied systems, these metrics should be sliced by microphone geometry, robot motion, ego-noise, reverberation, source overlap, device shift, compute, and real-time factor. *Semantic Perception* metrics test whether audio has been converted into action-relevant symbols. ASR should report WER or CER, but also entity WER, semantic preservation, and downstream intent or slot metrics when commands, names, numbers, or spatial relations matter [198], [199], [200]. Spoken-command and SLU settings use command accuracy, intent accuracy, and slot F1; captioning and audio-language grounding use BLEU, METEOR, CIDEr, SPICE, SPIDER, and audio-aware caption metrics such as MACE [201]. SUPERB, HEAR, AudioBench, and AIR-Bench further motivate reporting task-family scores separately across speech, paralinguistic, environmental, and music understanding [117], [131], [132], [133]. *Physical and Social Perception* metrics ask whether observable contact, material, speaker, and affective cues are reliable. Physical perception uses contact/event classification, temporal alignment, material or operation recognition, and generalization across surfaces, tools, and unseen objects. Social perception uses accuracy, unweighted average recall, macro-F1, arousal-valence MAE/RMSE, correlation, and concordance correlation coefficient, as in affective benchmarks such as AVEC 2019 [144]. Speaker-sensitive systems should additionally report verification or diarization errors, overlap-sensitive performance, and cross-speaker, cross-accent, cross-culture, and confidence-calibration slices.

Reason-stage metrics. Reason-stage metrics evaluate whether audio changes a belief, answer, plan, or explanation rather than merely producing a better perceptual label. *Acoustic Reasoning* is most developed in audio-visual navigation. It uses success rate, SPL, path length, distance to goal, collision rate, and generalization to unseen scenes or unheard sounds [202]; AVLEN adds diagnostics such as success weighted by number of actions, success when silent, distance to goal, and query behavior [23]. These metrics should always be paired with audio-off, corrupted-audio, delayed-audio, distractor-source, and oracle-audio variants so that improvement can be attributed to acoustic evidence rather than dataset priors. *Semantic Reasoning* metrics include exact match, answer accuracy, multiple-choice accuracy, token F1, compositional

subcategory scores, temporal/counting accuracy, abstention accuracy for unanswerable questions, and robustness to paraphrases or distractors. Open-ended AudioQA and AudioLLM evaluation should add evidence attribution, hallucination rate, confidence calibration, human preference, and calibrated LLM-as-judge or rubric scores. Benchmarks such as MMAU, MMAR, and the Audio Reasoning Challenge are useful because they separate final-answer correctness from multi-step reasoning and rationale quality [153], [158], [162]. *Physical Reasoning* metrics ask whether latent state estimates are correct: object or material classification, contact-state error, geometry or source-state error, count/temporal error for impacts, and physics-sensitive generation scores. PhyAVBench makes this idea explicit through audio-physics sensitivity tests and contrastive physical-response evaluation [169]. *Social Reasoning* metrics include sentiment or intent F1, user-state accuracy, arousal-valence error, turn-level consistency, sarcasm or intent scores, personalization benefit, fairness gaps, and uncertainty calibration. For embodied use, these metrics should be checked over interaction histories because a socially adaptive agent can be accurate on isolated clips but unstable, overconfident, or inappropriate across a session.

Interact-stage metrics. Interact-stage metrics measure the behavioral consequence of listening and reasoning. *Acoustic Interaction* asks whether movement improves perception, using separation gain, localization gain, target-to-interferer improvement, time-to-acquire, movement cost, collision or safety cost during active listening, and performance under moving sources, as in Move2Hear and active audio-visual separation [8], [9]. *Physical Interaction* metrics include task success, task completion rate, contact-state accuracy, failure-detection time, recovery rate, unsafe-contact rate, force or pose error, action latency, and policy change under audio removal or delay [17]. These metrics are strongest when they show that contact audio changes the action trajectory, not only a classifier output. *Semantic Interaction* metrics include clarification success, correction success, act/ask/refuse accuracy, turn-taking F1, endpointing latency, first-response latency, barge-in detection and recovery, repair turns, dialogue coherence, and explanation faithfulness. Dialogue evaluation work, including PARADISE-style combinations of task success, dialogue cost, and user satisfaction, shows why spoken interaction should not be judged by surface text overlap alone [203], [204]. *Social and Safety Interaction* metrics include trust, rapport, comfort, engagement, perceived intelligence, perceived safety, and satisfaction, with instruments such as the Godspeed questionnaire [205]. Safety diagnostics such as Succ-But-Unsafe and Violation Severity Index in SafeVLA-Bench [189], together with rollout failure signatures and runtime-cost diagnostics from Beyond Task Success [190], are useful templates for auditory agents. The central Interact-stage test is causal: removing, shuffling, corrupting, or delaying audio should reveal whether sound improves timing, recovery, safety, task completion, and human experience.

7 APPLICATIONS AND SYSTEMS

Auditory embodied intelligence is most useful when sound changes how an agent behaves in a concrete setting. This section therefore organizes applications by deployment scenario rather than by the three technical loops used in the preceding sections. We group current systems into industrial, educational, and service applications. These categories are not mutually exclusive: a factory assistant may need spoken dialogue, a classroom robot may need affective listening, and a home-care robot may need physical event detection. The point of this organization is to make the application demand explicit: what must be heard, what action should change, and what evidence would show that audio is causally useful. Fig. 8 summarizes the resulting application landscape.

7.1 Industrial Applications

Factory-floor perception and anomaly response. Industrial environments give audition a direct operational role because many safety and maintenance events are acoustic before they are visible. Mobile robots or fixed robot workcells can listen for alarms, abnormal machine vibration, air leaks, tool chatter, falling objects, collisions, and human calls for help. Embedded audition stacks such as ODAS [7] illustrate the low-level requirements: real-time localization, tracking, separation, and post-filtering under overlapping sources and reverberation. In smart manufacturing, these front ends must be integrated with visual inspection, proprioception, digital twins, and human-robot collaboration workflows [207], [208]. The application metric is not only event accuracy. A useful factory system should reduce time to detection, localize the source well enough for response, lower false alarms per shift, and trigger a safe robot or human action under high background noise.

Warehousing, packaging, and contact-rich manipulation. Industrial manipulation is a second major scenario because the robot’s own actions create informative sound. Packaging, sorting, insertion, wiping, pouring, fastening, and tool-use tasks produce contact signals that reveal slip, scraping, impact, empty containers, blocked motion, or failed engagement earlier than vision alone. Recent contact-audio systems show how interaction sounds can supervise dynamic manipulation, support audio-visual policy learning, and provide object-state evidence [15], [16], [60], [62], [63], [64]. Audio-VLA further shows how contact audio can be inserted into a vision-language-action policy rather than treated as a separate classifier [17]. For factory deployment, this line connects naturally to packaging and assembly pipelines, where recent VLA case studies report workflow failures and deployment constraints [222]. Strong evidence in this scenario requires audio-off, delayed-audio, and distractor-audio ablations, together with task success, recovery rate, unsafe-contact rate, and audio-to-action latency.

Inspection, maintenance, and collaborative safety. Inspection robots, warehouse robots, and collaborative manipulators also need auditory awareness of people and moving equipment. Speech and paralinguistic cues can signal task requests, urgency, warnings, and misunderstanding; nonverbal sound can communicate robot state, motion intent, or completion to nearby workers. Work on proactive collaboration and human-centric manufacturing emphasizes that the robot should be predictable and responsive, not merely accurate as a perception module [207], [209]. Designed robot sounds and sonification can make attention, waiting, uncertainty, or internal state more legible [18], [36]. Evaluation therefore includes near-miss reduction, interruption cost, worker response time, perceived safety, and trust calibration, in addition to standard speech or sound-event metrics.

7.2 Educational Applications

Classroom teaching assistants. In schools and universities, auditory embodied systems mainly support interaction rather than physical production. A classroom robot or teaching assistant must hear questions, answers, interruptions, laughter, hesitation, group participation, and signs of confusion while respecting student privacy. Robot-supported collaborative learning uses social robots as facilitators for small-group learning, where turn-taking, prompt timing, and balanced participation are central outcomes [210]. Pepper-like platforms and HRI turn-taking corpora illustrate the platform and interaction infrastructure for this setting [25], [211]. Audio is valuable when it helps the robot decide whether to invite a quieter student, ask a clarifying question, slow the pace, or hand control back to a teacher. Learning gain, participation

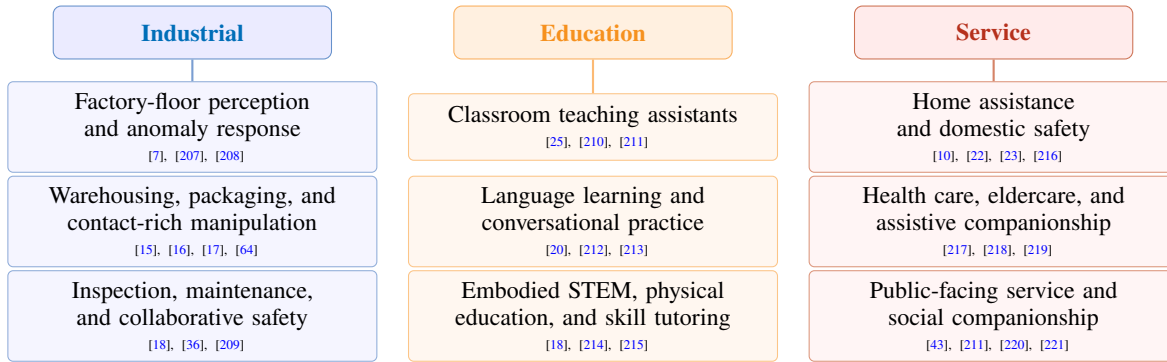


Fig. 8. Summary of auditory embodied applications.

balance, teacher workload, student comfort, and privacy protection are more meaningful than isolated recognition accuracy.

Language learning and conversational practice. Language education is a natural auditory application because the target skill itself is spoken interaction. A robot or embodied tutor can provide pronunciation feedback, conversational practice, listening exercises, prosody-sensitive correction, and adaptive dialogue. Studies of robot interaction styles for second-language conversation and AI-supported L2 education show the importance of engagement, feedback style, and behavioral participation [212], [213]. Speech-prosody grounding is also relevant because emphasis, hesitation, and adverbs can change how an instruction should be interpreted or practiced [20], [21]. In this scenario, audio should improve practice quality: more turns taken, better pronunciation or fluency over time, appropriate repair after ASR errors, and reduced learner anxiety. The system should avoid overfitting to accent, age, or language background and should expose uncertainty instead of correcting with unwarranted confidence.

Embodied STEM, physical education, and skill tutoring. Educational robots can also teach embodied skills: laboratory procedures, mathematics with manipulatives, music or rhythm activities, sports practice, and physical education. Here audition connects the learner’s action to feedback. Tapping, timing, tool contact, spoken self-explanation, and task sounds can reveal whether a learner is following a procedure or needs help. Work on embodied design for mathematics and voice-interactive educational robots points to this broader role of physical and vocal interaction in learning [214], [215]. Sonified feedback can also make invisible process states or robot uncertainty easier to perceive [18]. The key evaluation question is whether the auditory channel improves instruction timing, skill acquisition, retention, or learner agency, not whether the robot can label a sound clip in isolation.

7.3 Service Applications

Home assistance and domestic safety. Domestic service robots use audition for events that happen outside the camera view or before a user explicitly asks for help. Household systems may listen for smoke alarms, carbon-monoxide alarms, glass breaking, knocking, appliance events, running water, dropped objects, or spoken requests. Amazon Astro illustrates the deployed product form of sound-triggered domestic monitoring, where a mobile platform can investigate or notify a user after detecting safety-relevant sounds [216]. Audio-visual navigation and audio-language mapping show how a robot can move toward sounding targets, use sound as a spatial landmark, or answer navigation queries grounded in audio, vision, and language [10], [11], [22], [23]. The main deployment risks are privacy, false alarms, household reverberation,

overlapping speakers, and user control over when the robot listens. Metrics should include response latency, false alarms per day, localization quality, task completion, and user override behavior.

Health care, eldercare, and assistive companionship. Health and eldercare services combine safety monitoring, spoken dialogue, reminders, social support, and escalation to caregivers. ElliQ represents a deployed older-adult companion built around proactive conversation, reminders, and engagement [217]; ENRICHME studies perception and interaction for assistive robots in elderly homes [218]. Health-care-oriented HRI also includes robot-patient dialogue and older adult perceptions of robots as communication partners [219], [223]. In these applications, audio may provide check-ins, medication reminders, distress cues, conversation history, or evidence that a user did not respond. Because the stakes are high, audio inference should be conservative: the robot should distinguish social engagement from clinical judgment, maintain consent and data minimization, and escalate uncertain cases to humans rather than acting on a fragile affect or health estimate. Evaluation should include longitudinal engagement, comfort, privacy, escalation accuracy, missed-event rate, and caregiver workload.

Public-facing service and social companionship. Service robots in museums, hotels, restaurants, malls, airports, and public institutions depend on robust speech interaction, turn-taking, repair, and socially legible feedback. Pepper, Jibo, and Aibo show different platform forms for voice-centered social interaction and companionship [211], [220], [221]. Service encounters also expose social risks: users may feel embarrassed, ignored, interrupted, or over-monitored when a robot mishears them or responds with the wrong social tone [224], [225]. Vocal affect, hesitation, silence, and overlap can help a robot choose whether to repeat, clarify, apologize, simplify instructions, or yield to a human staff member [43]. Designed sounds can further communicate availability, waiting, completion, or warning, but they must be evaluated for annoyance, accessibility, and cultural interpretability [18], [36]. Across public service and companionship, the strongest auditory systems will be those that improve task completion and user comfort while making their listening behavior transparent and controllable.

Across these three application domains, the common lesson is that audio should be evaluated through its behavioral consequence. Industrial robots should detect and recover from physical events earlier; educational robots should improve participation and learning; service robots should provide safer, more natural, and more transparent assistance. Recognition scores remain useful diagnostics, but application evidence requires matched ablations, latency measurements, privacy analysis, and user- or task-level outcomes.

8 CHALLENGES AND FUTURE DIRECTIONS

The preceding sections show that auditory embodied intelligence is no longer a single perception problem. Sound must be extracted from real acoustic fields, grounded into physical and semantic state, and used to change robot action or human interaction. The open problems therefore appear at three coupled levels: the dataset level, where current resources rarely record complete closed loops; the stage level, where perception, reasoning, and interaction each expose different failure modes; and the system level, where audio must be integrated with vision, language, proprioception, safety, and long-term user experience. This section synthesizes these challenges using the same Physical Interaction Loop, Semantic Grounding-and-Action Loop, and Socio-Affective Interaction Loop used throughout the survey.

8.1 Dataset and Benchmark Bottlenecks

From audio datasets to embodied episodes. Current audio resources are strong for perceptual supervision but weak for embodied causality. Large-scale sound-event and audio-visual datasets such as AudioSet [81], VGGSound [87], and FSD50K [82] cover broad acoustic categories, and captioning or Audio Question Answering (AQA) resources such as AudioCaps [126], Clotho [127], WavCaps [128], Clotho-AQA [151], MMAU [153], PAQA [154], and DCASE 2025 AQA [155] support audio-language understanding. However, most of these datasets are clip-level, observer-centric, and offline rather than action-conditioned resources like embodied navigation, manipulation, or HRI protocols. They rarely include robot pose, action history, actuator state, microphone placement on a moving body, proprioception, recovery behavior, or human feedback. As a result, they can train a model to recognize what a sound is, but not whether that sound should make an embodied agent move, stop, ask, or refuse.

Missing embodied acoustics. Spatial and real-room datasets such as LOCATA [6], STARSS23 [90], CHiME-6 [96], DiPCo [97], DIRHA [98], ACE [107], OpenAIR [103], and MeshRIR [105] are essential for localization, tracking, reverberation, and far-field speech. Yet many resources still underrepresent robot-specific conditions: ego-noise from motors and wheels, changing microphone geometry, self-occlusion, wind and floor noise, asynchronous sensors, moving arrays, and non-stationary sources. Contact and physical-audio datasets are even narrower because existing resources focus on impacts, object identity, or material cues rather than the full closed-loop contact process. Greatest Hits [134], ObjectFolder [163], ObjectFolder 2.0 [164], and RealImpact [166] connect sound to materials, impacts, and multisensory objects, but they cover only a fraction of the contact modes needed for manipulation, such as sliding, pouring, scraping, insertion, tool use, breakage, and incipient failure.

Social data without social deployment. Socio-affective audio also suffers from a gap between labels and deployment. IEMOCAP [135] and RAVDESS [136] provide useful emotional speech supervision, while SEWA DB [182] and NoXi+J [184] add more naturalistic audio-visual interaction and cultural variation. Still, social-robot use requires longitudinal consent, privacy, personalization, relationship dynamics, and culturally situated norms. Short clips or acted labels cannot determine whether an adaptive robot response is appropriate, whether affect inference improves trust, or whether continuous listening creates discomfort. Future social datasets should therefore record not only audio and labels, but also interaction policies, user consent, uncertainty, adaptation decisions, and post-interaction outcomes.

Benchmarking the causal value of sound. The most important benchmark defect is the lack of matched closed-loop ablations that

isolate whether sound causally changes behavior. A mature auditory embodied benchmark should ask what changes when audio is removed, delayed, corrupted, shuffled across episodes, replaced by oracle signals, or attacked by distractor sources. Navigation benchmarks such as SoundSpaces [10], SoundSpaces 2.0 [11], AVLEN [23], and Sonicverse [14] show how acoustic simulation can define embodied tasks, while manipulation systems such as That Sounds Right [64], Mani-WAV [62], Hearing Touch [63], SonicSense [60], and Audio-VLA [17] show that contact audio can affect action. The next benchmark layer should make the embodied episode, rather than the isolated clip, the basic unit: synchronized audio, vision, proprioception, robot actions, dialogue turns, task state, human feedback, safety events, latency, recovery, and public protocols for audio ablation should be reported together.

8.2 Percept Stage

Physical Interaction Loop: robust robot audition under motion. The physical perceptual challenge is to listen while the robot is acting. Classical robot audition and systems such as ODAS [7] demonstrate real-time localization, tracking, separation, and post-filtering on embedded platforms, but embodied deployment adds stronger disturbances than static benchmarks: robot ego-noise, moving microphones, changing head or arm poses, occluded sources, reflections from nearby objects, and simultaneous speech or impact events. Future perceptual front ends should jointly estimate external sources and self-generated sound, expose calibrated uncertainty, and report performance by motion state, array geometry, acoustic condition, and real-time factor. Without such diagnostics, a high localization or separation score may not transfer to navigation, manipulation, or HRI.

Semantic Grounding-and-Action Loop: reverberant, multi-speaker, action-relevant speech. Semantic listening is not equivalent to generic ASR because robot language understanding must be grounded in action and environment state. A robot must preserve names, objects, spatial relations, negation, urgency, deictic expressions, and prosody under far-field, multi-speaker, and reverberant conditions. Spoken-command and SLU datasets such as Fluent Speech Commands [176] and Snips [177] are useful starting points, while Robots That Use Language [41] emphasizes that robot language understanding must be tied to action and environment state. The challenge is to evaluate semantic preservation rather than transcript quality alone: a word error is harmless if the plan is unchanged, but a small error in an object name, location, or safety constraint can be catastrophic. Future benchmarks should therefore score ASR, intent, slot, grounding, and downstream action jointly.

Socio-Affective Interaction Loop: privacy-preserving vocal state sensing. Social perception from voice is inherently uncertain and privacy-sensitive in embodied HRI. Affect and engagement resources such as IEMOCAP [135], CMU-MOSEI [171], MELD [172], and SEWA DB [182] support recognition of emotion, sentiment, or conversational state, but embodied agents must avoid overclaiming hidden mental states from noisy acoustic cues. The open problem is not only higher affect accuracy; it is calibrated, consent-aware, culturally robust inference that can decide when not to infer or when to ask. Promising directions include on-device feature extraction, privacy-preserving representations, confidence-aware adaptation, and evaluation protocols that measure comfort, trust, fairness gaps, and user control over audio data.

8.3 Reason Stage

Physical Interaction Loop: causal acoustic world models. Physical reasoning requires models that predict how actions, geometry, materials,

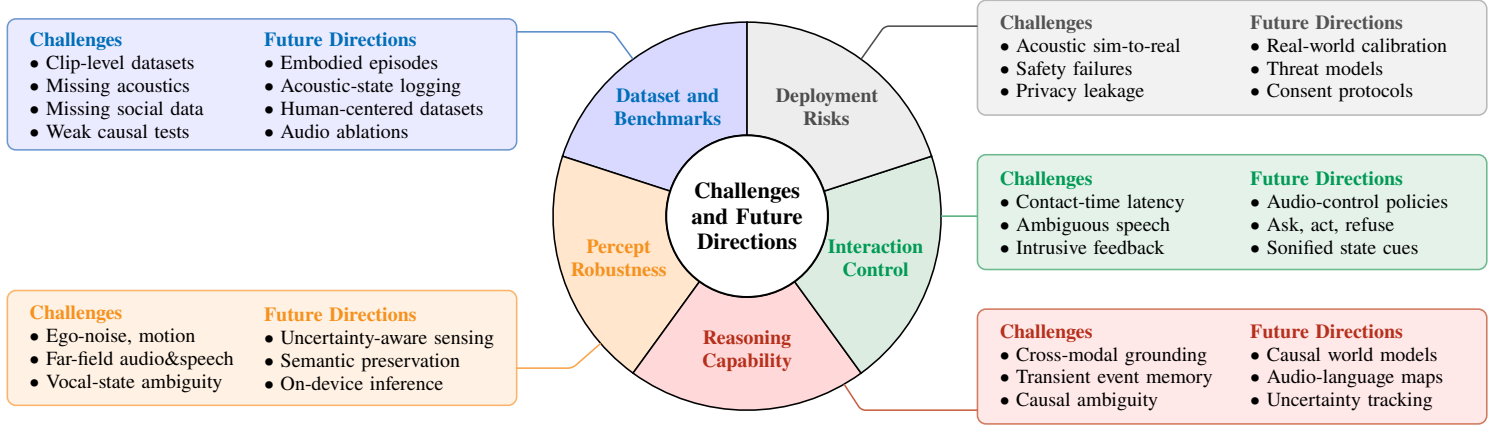


Fig. 9. Summary of open challenges and future directions in auditory-centered embodied intelligence. The central ring follows the structure of this section, while the surrounding panels highlight representative bottlenecks and future directions across datasets, perception, reasoning, interaction, and deployment.

contact, and robot motion change sound over time. SoundSpaces [10], SoundSpaces 2.0 [11], and Sonicverse [14] provide controlled spatialized environments, while physical-audio resources such as RealImpact [166] and ObjectFolder 2.0 [164] connect sound to objects, impacts, and materials. The reasoning challenge is causal rather than merely transferable: the agent must distinguish external events from self-generated sound, infer which hidden physical state produced an acoustic change, and predict how candidate actions will alter future sight-and-sound observations. Audio-Visual World Models [226] are a promising direction because they can represent these action-conditioned dynamics, but they should be evaluated with interventions that test physical causality rather than video-audio correlation alone. More general audio-aware foundation and VLA models should likewise treat audio as a first-class state channel, as illustrated by Audio-VLA [17] for contact-audio manipulation and NaVLA [24] for audio-conditioned navigation, while supporting streaming audio tokens, event memory, uncertainty estimates, and intervention tests that verify whether acoustic evidence actually changes action.

Semantic Grounding-and-Action Loop: persistent audio-language spatial memory. Semantic reasoning requires memory that binds sounds, words, objects, and places over time. AVLMaps [22] stores audio, vision, and language in spatial maps for robot navigation, AVLEN [23] and CAVEN [26] add query behavior and conversational help in noisy audio-visual navigation, and NaVLA [24] suggests a path toward unified vision-language-audio-action navigation. The challenge is persistence across delays, movements, and intermittent observations. A robot may hear a microwave beep, a person call from another room, or a dropped object outside the camera view, and then need to remember that event after several turns and movements. Future systems should maintain time-stamped audio-language memories with uncertainty, decay, source identity, and provenance, and should be evaluated on delayed grounding, contradiction resolution, and recovery after wrong acoustic hypotheses.

Socio-Affective Interaction Loop: long-term user-state and norm reasoning. Socio-affective reasoning must connect vocal cues to interaction policy across sessions. A Survey on Dialogue Management in Human-Robot Interaction [42] and Recent Advancements in Multimodal Human-Robot Interaction [43] show that social interaction depends on turn structure, task state, multimodal context, and norms rather than isolated utterances. The open challenge is to model user state without turning uncertain affect estimates into rigid labels. Long-term robots should distinguish transient acoustic states from stable preferences,

represent uncertainty explicitly, and adapt only when the benefit justifies the privacy and social risk. Evaluation should include longitudinal user studies, opt-in personalization, cultural and language variation, and failure cases such as misread distress, inappropriate empathy, or excessive questioning.

8.4 Interact Stage

Physical Interaction Loop: low-latency contact-audio control. Interaction-stage physical audio must support control at the timescale of contact. Systems such as That Sounds Right [64], Hearing Touch [63], ManiWAV [62], SonicSense [60], and Audio-VLA [17] show that interaction sounds and in-hand vibration can improve manipulation or object-state inference. The next challenge is to close the loop fast enough for recovery: detecting slip, collision, empty pouring, scraping, tool chatter, or unsafe impact before visual evidence is available. Future policies should fuse contact audio with force, touch, proprioception, and vision; report audio-to-action latency; and evaluate recovery rate, unsafe-contact rate, and task success under audio-off, delayed-audio, and distractor-audio conditions.

Semantic Grounding-and-Action Loop: deciding when to ask, act, or refuse. Spoken interaction becomes embodied when the robot must choose between acting, asking for clarification, refusing unsafe instructions, or explaining uncertainty. AVLEN [23] and CAVEN [26] make help-seeking part of navigation, while the HRI Turn-Taking Corpus [25] and dialogue-management work [42] show that timing and floor control affect whether spoken interaction succeeds. Future semantic-interaction systems should optimize act/ask/refuse decisions under acoustic uncertainty and task risk. Metrics should include clarification success, correction success, endpointing latency, repair cost, unsafe action prevention, and user burden, rather than only final task success or response fluency.

Socio-Affective Interaction Loop: streaming turn-taking and sonified feedback. For social robots, audio is both input and output. The robot must listen to overlapping speech, backchannels, silence, and prosody while also producing speech, nonverbal sounds, or sonified feedback that makes its internal state legible. Hearing the Robot’s Mind [18] and The Role of Consequential and Functional Sound in Human-Robot Interaction [36] show that designed sound can make robot state and intent more transparent, while Nonverbal Sound in Human-Robot Interaction [37] highlights the importance of comfort, interpretation, and social context. The challenge is to make feedback useful without

becoming intrusive. Future work should evaluate startle, annoyance, perceived safety, timing, interruption cost, and trust, and should test whether sound output improves coordination during real tasks rather than only subjective preference in isolated demos.

8.5 Deployment

Calibration and acoustic sim-to-real protocols. The acoustic sim-to-real gap is primarily a deployment challenge because models that work in controlled simulators can fail under new rooms, microphones, robot bodies, materials, motions, and distractor sources. Future deployment protocols should pair physically grounded simulators with real-world calibration sets, similar in spirit to room-acoustic and impact-field resources. A policy trained in simulation should be tested under shifted acoustics, changed microphone placement, new objects, new robot motions, and adversarial or distracting sounds. Progress should be reported not only by success rate but also by how performance degrades as the acoustic gap increases.

Safety, privacy, and threat models. Auditory systems are vulnerable to false alarms, spoofing, speech injection, deepfake voices, privacy leakage, and unsafe action triggered by misleading sounds. Diagnostic benchmarks such as SafeVLA-Bench [189] and Beyond Task Success [190] provide useful templates for separating task success from safety, latency, and failure signatures, but auditory agents require additional threat models: hidden speakers, replayed commands, inaudible or ultrasonic interference, private conversation leakage, and adversarial environmental sounds. Future evaluation should report refusal precision and recall, attack success rate, privacy leakage, calibration under uncertainty, and safe recovery after wrong acoustic beliefs.

Long-term human-centered deployment. Finally, embodied audio must be studied over time because household monitoring, eldercare, collaborative work, and companion robots involve repeated interaction, changing users, changing rooms, and evolving privacy expectations. Social platforms and studies such as Pepper [211], Jibo [220], Aibo [221], ElliQ [217], and sonified HRI [18] illustrate the demand for auditory interaction, but future work should move from short demonstrations to longitudinal protocols. The core research question is whether audio-aware adaptation improves task performance, trust, independence, and comfort without creating surveillance pressure, dependency, or culturally inappropriate behavior.

9 CONCLUSION

This work provides a structured review of multimodal auditory learning for embodied intelligence. We introduce a two-axis loop-stage taxonomy whose loop axis comprises the Physical Interaction Loop, the Semantic Grounding-and-Action Loop, and the Socio-Affective Interaction Loop, while its stage axis comprises Percept, Reason, and Interact. Using this taxonomy, we organize and compare how existing approaches acquire and represent auditory information, integrate it with other modalities, and use it to support embodied behavior. We further summarize representative methods, datasets, benchmarks, and evaluation metrics, and discuss current applications, challenges, and future directions.

Collectively, the reviewed literature indicates an ongoing but incomplete transition from treating audition as an isolated perceptual capability to incorporating auditory information as situated evidence for embodied reasoning and behavior. Although numerous studies report improvements in recognition, localization, and multimodal representation, comparatively few establish that these perceptual advances translate

into reliable changes in an agent’s internal state estimation, decision-making, action selection, or interaction strategy. Assessing this relationship is further complicated by heterogeneous evaluation protocols, which frequently emphasize task-specific performance without isolating the contribution of auditory information to downstream behavioral outcomes. Advancing the field will therefore require more systematic integration of auditory perception with reasoning and control, evaluation under realistic acoustic and embodied conditions, and matched ablation studies that characterize when, how, and to what extent audition improves agent behavior.

REFERENCES

- [1] M. Soori, B. Arezoo, and R. Dastres, “Artificial intelligence, machine learning and deep learning in advanced robotics, a review,” *Cognitive Robotics*, vol. 3, pp. 54–70, 2023.
- [2] Y. Kim, D. Kim, J. Choi, J. Park, N. Oh, and D. Park, “A survey on integration of large language models with intelligent robots,” *Intelligent Service Robotics*, vol. 17, no. 5, pp. 1091–1107, 2024.
- [3] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, J. Tompson, Q. Vuong, T. Yu, W. Huang *et al.*, “PaLM-E: An embodied multimodal language model,” in *ICML*, 2023.
- [4] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, T. Ding, D. Driess, C. Finn, P. Florence, C. Fu *et al.*, “RT-2: Vision-language-action models transfer web knowledge to robotic control,” in *CoRL*, 2023.
- [5] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, “OpenVLA: An open-source vision-language-action model,” in *CoRL*, 2024.
- [6] C. Evers, H. W. Lollmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, “The locata challenge: Acoustic source localization and tracking,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.
- [7] F. Grondin, D. Létourneau, C. Godin, J.-S. Lauzon, J. Vincent, S. Michaud, S. Faucher, and F. Michaud, “Odas: Open embedded audition system,” *Frontiers in Robotics and AI*, vol. 9, p. 854444, 2022.
- [8] S. Majumder, Z. Al-Halah, and K. Grauman, “Move2hear: Active audio-visual source separation,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 275–285.
- [9] S. Majumder and K. Grauman, “Active audio-visual separation of dynamic sound sources,” in *Lecture Notes in Computer Science*, 2022, pp. 551–569.
- [10] 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.
- [11] C. Chen, C. Schissler, S. Garg, P. Kobernik, A. Clegg, P. Calamia, D. Batra, P. Robinson, and K. Grauman, “Soundspaces 2.0: A simulation platform for visual-acoustic learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 8896–8911, 2022.
- [12] C. Chen, S. Majumder, Z. Al-Halah, R. Gao, S. K. Ramakrishnan, and K. Grauman, “Learning to set waypoints for audio-visual navigation,” in *ICLR*, 2021.
- [13] S. Purushwalkam, S. V. A. Gari, V. K. Ithapu, C. Schissler, P. Robinson, A. Gupta, and K. Grauman, “Audio-visual floorplan reconstruction,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1163–1172.
- [14] R. Gao, H. Li, G. Dharan, Z. Wang, C. Li, F. Xia, S. Savarese, L. Fei-Fei, and J. Wu, “Sonicverse: A multisensory simulation platform for embodied household agents that see and hear,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 704–711.
- [15] 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*, 2022.
- [16] 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.
- [17] X. Wei, H. Zhang, X. Cao, S. Xie, W. Ge, Y. Li, and C. Wang, “Audio-vla: Adding contact audio perception to vision-language-action model for robotic manipulation,” *arXiv preprint arXiv:2511.09958*, 2025.
- [18] S. Arreghini, A. Paolillo, G. Abbate, and A. Giusti, “Hearing the robot’s mind: sonification for explicit feedback in human-robot interaction,” in *International Workshop on Human-Friendly Robotics*. Springer, 2024, pp. 45–57.

- [19] W. Zhao, P. Ding, Z. Min, Z. Gong, S. Bai, H. Zhao, and D. Wang, "Vlas: Vision-language-action model with speech instructions for customized robot manipulation," in *International conference on learning representations*, vol. 2025, 2025, pp. 51 676–51 693.
- [20] D. Sasu, B. Quartey, K. A. Yamoah, and N. Schluter, "Enhancing speech instruction understanding and disambiguation in robotics via speech prosody," in *Interspeech 2025*, 2025, pp. 1958–1962.
- [21] F. Lotfi, F. Faraji, N. Kakodkar, T. Manderson, D. Meger, and G. Dudek, "Constrained robotic navigation on preferred terrains using llms and speech instruction: Exploiting the power of adverbs," in *International Symposium on Experimental Robotics*. Springer, 2023, pp. 118–128.
- [22] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Audio visual language maps for robot navigation," in *ISER*, 2023.
- [23] S. Paul, A. Roy-Chowdhury, and A. Cherian, "Avlen: Audio-visual-language embodied navigation in 3d environments," in *Advances in Neural Information Processing Systems 35*, 2022, pp. 6236–6249.
- [24] J. Fan, P. Chen, C. Li, Q. Du, J. Chen, and M. Tan, "Navla²: A vision-language-audio-action model for multimodal instruction navigation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 22, pp. 18 234–18 242, 2026.
- [25] J. Yang, P. Wang, Y. Zhu, M. Feng, M. Chen, and X. He, "Gated multimodal fusion with contrastive learning for turn-taking prediction in human-robot dialogue," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7747–7751.
- [26] X. Liu, S. Paul, M. Chatterjee, and A. Cherian, "Caven: An embodied conversational agent for efficient audio-visual navigation in noisy environments," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 4, 2024, pp. 3765–3773.
- [27] Y. Xu, X. Zhan, A. Y. Kaltungo, M. S. Ng, T. Ishizawa, K. Fujimoto, and C. Cheung, "Dialogue based interactive explanations for safety decisions in human robot collaboration," *arXiv preprint arXiv:2604.05896*, 2026.
- [28] R. Mishra, A. Frye, M. M. Rayguru, and D. O. Popa, "Personalized speech emotion recognition in human-robot interaction using vision transformers," *IEEE Robotics and Automation Letters*, vol. 10, no. 5, pp. 4890–4897, 2025.
- [29] S. Yang, X. Li, X. Fei, M. Li, and M. Li, "Bridging speech, emotion, and motion: a vlm-based multimodal edge-deployable framework for humanoid robots," *arXiv preprint arXiv:2602.07434*, 2026.
- [30] E. C. Montiel-Vazquez, C. A. Cruz, S. Gkikas, T. Kassiotis, G. Giannakakis, and R. Gomez, "Efficient emotion-aware iconic gesture prediction for robot co-speech," *arXiv preprint arXiv:2604.11417*, 2026.
- [31] C. A. Cruz, E. C. Montiel-Vazquez, C. Maeda, and R. Gomez, "When and how to express empathy in human-robot interaction scenarios," in *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2025, pp. 1070–1077.
- [32] I. Pedersen and A. Slane, "Better than "better than nothing": Design strategies for enculturated empathetic ai robot companions for older adults," *arXiv preprint arXiv:2510.01192*, 2025.
- [33] A. Nardelli, A. Sgorbissa, and C. T. Recchiuto, "Designing empathetic companions: Exploring personality, emotion, and trust in social robots," *arXiv preprint arXiv:2504.13964*, 2025.
- [34] Y. Meng, G. Fan, B. Liu, Y. Sun, R. Chen, and H. Mi, "Engagement is not transfer: A withdrawal study of a consumer social robot with autistic children at home," in *Proceedings of the 25th Annual ACM Interaction Design and Children Conference*, 2026, pp. 755–783.
- [35] G. Zhang, "Semantic co-speech gesture synthesis and real-time control for humanoid robots," *arXiv preprint arXiv:2512.17183*, 2025.
- [36] A. Smith and M. Kennedy, "The role of consequential and functional sound in human-robot interaction: Toward audio augmented reality interfaces," *arXiv preprint arXiv:2511.15956*, 2025.
- [37] B. J. Zhang and N. T. Fitter, "Nonverbal sound in human-robot interaction: A systematic review," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 4, pp. 1–46, 2023.
- [38] S. Argentieri, P. Danes, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech & Language*, vol. 34, no. 1, pp. 87–112, 2015.
- [39] R. Jalayer, M. Jalayer, and A. Baniasadi, "A review on sound source localization in robotics: Focusing on deep learning methods," *Applied Sciences*, vol. 15, no. 17, p. 9354, 2025.
- [40] M. Saveriano, F. J. Abu-Dakka, A. Kramberger, and L. Peternel, "Dynamic movement primitives in robotics: A tutorial survey," *The International Journal of Robotics Research*, vol. 42, no. 13, pp. 1133–1184, 2023.
- [41] S. Tellex, N. Gopalan, H. Kress-Gazit, and C. Matuszek, "Robots that use language: A survey," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 25–55, 2020.
- [42] M. M. Reimann, F. A. Kunneman, C. Oertel, and K. V. Hindriks, "A survey on dialogue management in human-robot interaction," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 2, pp. 1–22, 2024.
- [43] H. Su, W. Qi, J. Chen, C. Yang, J. Sandoval, and M. A. Laribi, "Recent advancements in multimodal human-robot interaction," *Frontiers in Neurorobotics*, vol. 17, 2023.
- [44] A. Howcroft, M. E. Giannaccini, S. Benford, A. Khan, and H. Blake, "Speech-touch integration for affective human-robot interaction: a scoping review," *Frontiers in Robotics and AI*, vol. 13, 2026.
- [45] E. Pashevich, "Can communication with social robots influence how children develop empathy—best-evidence synthesis," *AI & SOCIETY*, vol. 37, no. 2, pp. 579–589, 2022.
- [46] M. Leineweber, C. V. Keusgen, M. Bubeck, R. Ranisch, J. Haltaufderheide, and C. Klingler, "Ethical aspects of the use of social robots in caring for older people – a systematic qualitative review," *Medicine, Health Care and Philosophy*, vol. 29, no. 1, pp. 209–224, 2026.
- [47] V. Liu, T. Du, J. Sehn, J. Collier, and F. Grondin, "Sound source localization for human-robot interaction in outdoor environments," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 6121–6126.
- [48] V. Tourbabin and B. Rafaely, "Theoretical framework for the optimization of microphone array configuration for humanoid robot audition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1803–1814, 2014.
- [49] Y. Hong, M. Wang, Y. Liu, Y. Fu, K. Hung, and B. Tao, "Sky-ear: An unmanned aerial vehicle-enabled victim sound detection and localization system," *arXiv preprint arXiv:2604.12455*, 2026.
- [50] J. Ortigoso-Narro, J. A. Belloch, A. Amor-Martin, S. Roger, and M. Cobos, "Real-time object tracking with on-device deep learning for adaptive beamforming in dynamic acoustic environments," *The Journal of Supercomputing*, vol. 82, no. 3, 2026.
- [51] K. Nakada, K. Hoshiba, B. Yen, M. Kumon, and Y. Sasaki, "Swarm active audition with robots and drones: Real-world performance validation," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 6107–6112.
- [52] Z. Turcotte and F. Grondin, "Lend me an ear: Speech enhancement using a robotic arm with a microphone array," *arXiv preprint arXiv:2602.17818*, 2026.
- [53] L. Macesanu, B. Folefack, S. Singh, R. Ray, B. Abbatematteo, and R. Martin-Martin, "Caver: Curious audiovisual exploring robot," *arXiv preprint arXiv:2511.07619*, 2025.
- [54] H. Wang, Y. Wang, F. Zhong, M. Wu, J. Zhang, Y. Wang, and H. Dong, "Learning semantic-agnostic and spatial-aware representation for generalizable visual-audio navigation," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3900–3907, 2023.
- [55] G. Tatiya, J. Francis, L. Bondi, I. Navarro, E. Nyberg, J. Sinapov, and J. Oh, "Knowledge-driven scene priors for semantic audio-visual embodied navigation," *arXiv preprint arXiv:2212.11345*, 2022.
- [56] J. Li and Y. Yu, "Generalizable audio-visual navigation via binaural difference attention and action transition prediction," *arXiv preprint arXiv:2604.05007*, 2026.
- [57] T. Liu and Y. Yu, "Reliability-aware geometric fusion for robust audio-visual navigation," *arXiv preprint arXiv:2604.02391*, 2026.
- [58] Y. Yu, W. Huang, F. Sun, C. Chen, Y. Wang, and X. Liu, "Sound adversarial audio-visual navigation," in *International Conference on Learning Representations*, 2022.
- [59] C. Chen, J. Ramos, A. Tomar, and K. Grauman, "Sim2real transfer for audio-visual navigation with frequency-adaptive acoustic field prediction," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 8595–8602.
- [60] J. Liu and B. Chen, "Sonicsense: Object perception from in-hand acoustic vibration," in *Proceedings of The 8th Conference on Robot Learning*, vol. 270. PMLR, 06–09 Nov 2025, pp. 4332–4353.
- [61] H. Ju, S. Huang, H. Li, Z. Ding, S. Liu, M. Wang, and Z. Zheng, "From instruction to event: Sound-triggered mobile manipulation," *arXiv preprint arXiv:2601.21667*, 2026.
- [62] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song, "Maniwave: Learning robot manipulation from in-the-wild audio-visual data," *Conference on Robot Learning*, 2024.
- [63] J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, "Hearing touch: Audio-visual pretraining for contact-rich manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6912–6919.
- [64] A. Thankaraj and L. Pinto, "That sounds right: Auditory self-supervision for dynamic robot manipulation," in *CoRL*, 2023.
- [65] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [66] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "Slurp: A spoken language understanding resource package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7252–7262.

- [67] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Interspeech 2019*, 2019, pp. 814–818.
- [68] T. Kollar, S. Tellex, M. Walter, A. Huang, A. Bachrach, S. Hemachandra, E. Brunskill, A. Banerjee, D. Roy, S. Teller *et al.*, "Generalized grounding graphs: A probabilistic framework for understanding grounded commands," *arXiv preprint arXiv:1712.01097*, 2017.
- [69] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, pp. 1507–1514, 2011.
- [70] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sunderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3674–3683.
- [71] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 737–10 746.
- [72] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning*, 2022.
- [73] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," *Conference on Robot Learning*, 2022.
- [74] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur, and D. Hakkani-Tur, "Teach: Task-driven embodied agents that chat," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, pp. 2017–2025, 2022.
- [75] X. Gao, Q. Gao, R. Gong, K. Lin, G. Thattai, and G. S. Sukhatme, "DialFRED: Dialogue-enabled agents for embodied instruction following," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 049–10 056, 2022.
- [76] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur, and D. Hakkani-Tur, "Dialog acts for task driven embodied agents," in *SIGDIAL*, 2022.
- [77] V. Belcamino, M. Kilina, A. Carfi, V. Seidita, F. Mastrogiovanni, and A. Chella, "Factored reasoning with inner speech and persistent memory for evidence-grounded human-robot interaction," *arXiv preprint arXiv:2602.00675*, 2026.
- [78] Q. Ren, R. Proesmans, Y. Hou, F. wyffels, and T. Belpaeme, "Touch and tell: Multimodal decoding of human emotions and social gestures for robots," *arXiv preprint arXiv:2412.03300*, 2024.
- [79] G. Laban, J. Wang, and H. Gunes, "A robot-led intervention for emotion regulation: From expression to reappraisal," *IEEE Transactions on Affective Computing*, 2026.
- [80] Y. Zhu, L. Li, I. Qian, W. Zhou, Y. Yuan, Q. Li, N. Liu, and J. Zhang, "Awakening facial emotional expressions in human-robot," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 21 511–21 518.
- [81] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.
- [82] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.
- [83] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *ACM MM*, 2015.
- [84] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *ACM MM*, 2014.
- [85] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1128–1132.
- [86] I. Mohino-Herranz, J. García-Gómez, M. Aguilar-Ortega, M. Utrilla-Manso, R. Gil-Pita, and M. Rosa-Zurera, "Introducing the realised dataset for sound event classification," *Electronics*, vol. 11, no. 12, p. 1811, 2022.
- [87] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 721–725.
- [88] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, "Localizing visual sounds the hard way," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 862–16 871.
- [89] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, Nov. 2020, pp. 165–169.
- [90] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, "Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Advances in Neural Information Processing Systems 36*, 2023, pp. 72 931–72 957.
- [91] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "Dregon: Dataset and methods for uav-embedded sound source localization," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–8.
- [92] G. Jekateryńczuk, R. Szadkowski, and Z. Piotrowski, "Uavirbase: A public-access unmanned aerial vehicle sound source localization dataset," *Applied Sciences*, vol. 15, no. 10, p. 5378, 2025.
- [93] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 74–79.
- [94] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16.3: An audio-visual corpus for speaker localization and tracking," in *Lecture Notes in Computer Science*, 2005, pp. 182–195.
- [95] Inria Perception Team, "NAL and NAR: NAO audio localization and robot audition datasets," Dataset release, 2012.
- [96] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.
- [97] M. V. Segbroeck, A. Zaid, K. Kutsenko, C. Huerta, T. Nguyen, X. Luo, B. Hoffmeister, J. Trmal, M. Omologo, and R. Maas, "Dipco — dinner party corpus," in *Interspeech 2020*, 2020, pp. 434–436.
- [98] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The dirha-english corpus and related tasks for distant-speech recognition in domestic environments," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 275–282.
- [99] N. Bertin, E. Camberlein, R. Lebarbenchon, E. Vincent, S. Sivasankaran, I. Illina, and F. Bimbot, "Voicehome-2, an extended corpus for multichannel speech processing in real homes," *Speech Communication*, vol. 106, pp. 68–78, 2019.
- [100] B. Yang, C. Quan, Y. Wang, P. Wang, Y. Yang, Y. Fang, N. Shao, H. Bu, X. Xu, and X. Li, "Realman: A real-recorded and annotated microphone array dataset for dynamic speech enhancement and localization," in *Advances in Neural Information Processing Systems 37*, 2024, pp. 105 997–106 019.
- [101] R. M. Corey, M. D. Skarha, and A. C. Singer, "Massive distributed microphone array dataset," Illinois Data Bank, University of Illinois Urbana-Champaign, 2019.
- [102] D. Di Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, and S. Gannot, "dEchorate: A calibrated room impulse response dataset for echo-aware signal processing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, p. 39, 2021.
- [103] D. T. Murphy, "OpenAIR: The open acoustic impulse response library," OpenAIR website, 2010.
- [104] Institute of Communication Systems, RWTH Aachen University, "Aachen impulse response database," Dataset, Version 1.4, Oct. 2012.
- [105] S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, and J. Brunnström, "MeshRIR: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 1–5.
- [106] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.
- [107] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.
- [108] A. Kujawski, A. J. R. Pelling, and E. Sarraj, "MIRACLE—a microphone array impulse response dataset for acoustic learning," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 32, 2024.
- [109] A. J. R. Pelling, A. Kujawski, and E. Sarraj, "SRIRACHA: Shoebox Room Impulse Response Archive with Varying Absorption," Jul. 2025.
- [110] L. Friede, C. Tuna, F. Knauff, A. Prinn, K. Ersinadim, and A. Walther, "Multi-Purpose Room Impulse Response Dataset Measured on a 3D Spatial Grid," in *Proceedings of the 156th Audio Engineering Society Convention*. Audio Engineering Society, 2024, convention Paper 10702.
- [111] Z. Chen, I. D. Gebru, C. Richardt, A. Kumar, W. Laney, A. Owens, and A. Richard, "Real Acoustic Fields: An audio-visual room acoustics dataset and benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2024, pp. 21 886–21 896.
- [112] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The cipc hrtf database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, 2001, pp. 99–102.

- [113] R. Peterson, A. Tanelus, C. Ick, B. Mimica, N. Francis, V. Ivan, A. Choudhri, A. Falkner, M. Murthy, D. Schneider, D. Sanes, and A. Williams, "Vocal call locator benchmark (vcl) for localizing rodent vocalizations from multi-channel audio," in *Advances in Neural Information Processing Systems* 37, 2024, pp. 106 370–106 382.
- [114] M. Fuentes, B. Steers, P. Zinemanas, M. Rocamora, L. Bondi, J. Wilkins, Q. Shi, Y. Hou, S. Das, X. Serra, and J. P. Bello, "Urban sound & sight: Dataset and benchmark for audio-visual urban scene understanding," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 141–145. [Online]. Available: <https://doi.org/10.1109/icassp43922.2022.9747644>
- [115] Y. Piadyk, J. Rulff, E. Brewer, M. Hosseini, K. Ozbay, M. Sankaradas, S. Chakradhar, and C. Silva, "Streetaware: A high-resolution synchronized multimodal urban scene dataset," *Sensors*, vol. 23, no. 7, p. 3710, 2023.
- [116] A. Brunetto, S. Hornauer, S. X. Yu, and F. Moutarde, "The audio-visual batvision dataset for research on sight and sound," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 1–8.
- [117] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally, M. Henry, N. Pinto, C. Noufi, C. Clough, D. Herremans, E. Fonseca, J. Engel, J. Salamon, P. Esling, P. Manocha, S. Watanabe, Z. Jin, and Y. Bisk, "HEAR: Holistic evaluation of audio representations," in *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, ser. Proceedings of Machine Learning Research, vol. 176. PMLR, 2022, pp. 125–145.
- [118] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.
- [119] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [120] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.
- [121] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.
- [122] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, 2020.
- [123] A. Rousseau, P. Deléglise, and Y. Estève, "Ted-lium: an automatic speech recognition dedicated corpus," in *Proceedings of the Language Resources and Evaluation Conference*, 2012, pp. 125–129.
- [124] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23–26, 1992*, 1992.
- [125] D. Wang and X. Zhang, "Thchs-30 : A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015.
- [126] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *NAACL*, 2019.
- [127] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.
- [128] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3339–3354, 2024.
- [129] I. Martín-Morato and A. Mesaros, "MACS: Multi-annotator captioned soundscapes," Zenodo, 2021.
- [130] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [131] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhota, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "Superb: Speech processing universal performance benchmark," in *Interspeech 2021*, 2021, pp. 1194–1198.
- [132] B. Wang, X. Zou, G. Lin, S. Sun, Z. Liu, W. Zhang, Z. Liu, A. Aw, and N. F. Chen, "Audiobench: A universal benchmark for audio large language models," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 4297–4316.
- [133] Q. Yang, J. Xu, W. Liu, Y. Chu, Z. Jiang, X. Zhou, Y. Leng, Y. Lv, Z. Zhao, C. Zhou, and J. Zhou, "Air-bench: Benchmarking large audio-language models via generative comprehension," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 1979–1998.
- [134] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2405–2413.
- [135] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [136] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, p. e0196391, 2018.
- [137] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.
- [138] F. Burkhardt, A. Paeschke, M. Rolles, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Interspeech 2005*, 2005, pp. 1517–1520.
- [139] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017*, 2017, pp. 2616–2620.
- [140] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech 2018*, 2018, pp. 1086–1090.
- [141] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *Interspeech 2016*, 2016, pp. 818–822.
- [142] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit," University of Edinburgh, Centre for Speech Technology Research, 2019, version 0.92, dataset.
- [143] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A non-native english speech corpus," in *INTERSPEECH*, 2018.
- [144] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallo-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 3–12.
- [145] Z. Shi, L. Zhang, L. Li, and Y. Shen, "Towards audio-visual navigation in noisy environments: A large-scale benchmark dataset and an architecture considering multiple sound-sources," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 14, pp. 14 673–14 680, 2025.
- [146] H. Ryu, J. S. Chung, and D. Harwath, "Hear you are: Teaching LLMs spatial reasoning with vision and spatial sound," in *CVPR*, 2026.
- [147] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 186–190.
- [148] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in *CVPR*, 2021, pp. 15 511–15 520.
- [149] Y. Wang, Y. Yu, F. Sun, L. Wang, and W. Zheng, "Audio-guided visual perception for audio-visual navigation," in *2025 International Conference on Virtual Reality and Visualization (ICVRV)*. IEEE, Dec. 2025.
- [150] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 19 086–19 096.
- [151] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1140–1144.
- [152] H. Alamri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson, S. Lee, and D. Parikh, "Audio visual scene-aware dialog," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7550–7559.
- [153] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "MMAU: A massive multi-task audio understanding and reasoning benchmark," in *International Conference on Learning Representations*, 2025, pp. 84 929–84 964.
- [154] J. Wang, Y. Niu, D. Xu, and Z. Wei, "Listen, pause, and reason: Toward perception-grounded hybrid reasoning for audio understanding," in *Findings of the Association for Computational Linguistics: ACL 2026*, 2026, pp. 35 653–35 671.
- [155] C.-H. H. Yang, S. Ghosh, Q. Wang, J. Kim, H. Hong, S. Kumar, G. Zhong, Z. Kong, S. Sakshi, V. Lokegaonkar, O. Nieto, R. Duraiswami, D. Manocha, G. Kim, J. Du, R. Valle, and B. Catanzaro, "Multi-domain audio question answering benchmark toward acoustic content reasoning," in *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026.

- [156] S. Ghosh, A. Seth, S. Kumar, U. Tyagi, C. K. R. Evuru, R. S. S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "CompA: Addressing the gap in compositional reasoning in audio-language models," in *The Twelfth International Conference on Learning Representations*, 2024.
- [157] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 6288–6313.
- [158] Z. Ma, Y. Ma, Y. Zhu, C. Yang, Y.-W. Chao, R. Xu, W. Chen, Y. Chen, Z. Chen, J. Cong *et al.*, "MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix," in *Advances in Neural Information Processing Systems*, vol. 38, 2025.
- [159] Y. Li, G. Zhang, Y. Ma, R. Yuan, Zhu, H. Guo, Y. Liang, J. Liu, N. Wang, J. Yang, S. Wu, X. Qu, J. Shi, X. Zhang, Z. Yang, Y. WEN, Y. Wang, S. Li, Z.-X. ZHANG, R. Liu, E. Benetos, W. Huang, and C. Lin, "Omnibench: Towards the future of universal omni-language models," in *Advances in Neural Information Processing Systems*, vol. 38, 2025.
- [160] D. Wang, J. Li, J. Wu, D. Yang, X. Chen, T. Zhang, and H. Meng, "Mmsu: A massive multi-task spoken language understanding and reasoning benchmark," *arXiv preprint arXiv:2506.04779*, 2025.
- [161] Z. Sun, S. Wang, Z. Lin, C. Wang, D. Gao, Y. Cao, C. He, P. Zhou, and L. Xie, "Msu-bench: Towards speaker-centric understanding in conversational multi-speaker scenarios," *arXiv preprint arXiv:2606.22868*, 2026.
- [162] Z. Ma, R. Xu, Y. Ma, C.-H. H. Yang, B. Li, J. Kim, J. Xu, J. Li, C. Busso, K. Yu *et al.*, "The interspeech 2026 audio reasoning challenge: Evaluating reasoning process quality for audio reasoning models and agents," *arXiv preprint arXiv:2602.14224*, 2026.
- [163] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei, and J. Wu, "Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations," *arXiv preprint arXiv:2109.07991*, 2021.
- [164] R. Gao, Z. Si, Y.-Y. Chang, S. Clarke, J. Bohg, L. Fei-Fei, W. Yuan, and J. Wu, "Objectfolder 2.0: A multisensory object dataset for sim2real transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 598–10 608.
- [165] R. Gao, Y. Dou, H. Li, T. Agarwal, J. Bohg, Y. Li, L. Fei-Fei, and J. Wu, "The objectfolder benchmark: Multisensory learning with neural and real objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 17 276–17 286.
- [166] S. Clarke, R. Gao, M. Wang, M. Rau, J. Xu, J.-H. Wang, D. L. James, and J. Wu, "Realimpact: A dataset of impact sound fields for real objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1516–1525.
- [167] A. Sterling, J. Wilson, S. Lowe, and M. C. Lin, "Isnn: Impact sound neural network for audio-visual object classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 555–572.
- [168] C. Gan, Y. Gu, S. Zhou, J. Schwartz, S. Alter, J. Traer, D. Gutfreund, J. B. Tenenbaum, J. H. McDermott, and A. Torralba, "Finding fallen objects via asynchronous audio-visual integration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 523–10 533.
- [169] T. Xie, W. Lei, K. Jiang, G. Huang, P. Zhang, C. Zhang, F. Ma, H. He, H. Zhang, J. He *et al.*, "Phyabench: A challenging audio physics-sensitivity benchmark for physically grounded text-to-audio-video generation," *arXiv preprint arXiv:2512.23994*, 2025.
- [170] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.
- [171] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
- [172] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 527–536.
- [173] J. Zhao, T. Zhang, J. Hu, Y. Liu, Q. Jin, X. Wang, and H. Li, "M3ed: Multimodal multi-scene multi-label emotional dialogue database," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5699–5710.
- [174] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an Obviously perfect paper)," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 4619–4629.
- [175] H. Zhang, H. Xu, X. Wang, Q. Zhou, S. Zhao, and J. Teng, "Mintrec: A new dataset for multimodal intent recognition," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 1688–1697.
- [176] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Interspeech 2019*, 2019, pp. 814–818.
- [177] A. Saade, J. Dureau, D. Leroy, F. Caltagirone, A. Coucke, A. Ball, C. Doumouro, T. Lavril, A. Caulier, T. Bluche *et al.*, "Spoken language understanding on the edge," in *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, 2019, pp. 57–61.
- [178] H. Liang, S. Li, X. Ma, N. Hendrich, T. Gerkmann, F. Sun, and J. Zhang, "Making sense of audio vibration for liquid height estimation in robotic pouring," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 5333–5339.
- [179] A. Sawhney, S. Lee, K. Zhang, M. Veloso, and O. Kroemer, "Playing with food: Learning food item representations through interactive exploration," in *International Symposium on Experimental Robotics*. Springer, 2020, pp. 309–322.
- [180] K. M. Rawf and A. Abdulrahman, "Multi-keyboard acoustic (mka) datasets," Mendeley Data, Version 4, 2024.
- [181] A. Barahona-Ríos and S. Pauletto, "Knocking sound effects with emotional intentions," Zenodo, 2020.
- [182] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 1022–1040, 2019.
- [183] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, vol. 7, no. 1, p. 293, 2020.
- [184] M. Funk, S. Okada, and E. André, "Multilingual dyadic interaction corpus noxi+j: Toward understanding asian-european non-verbal cultural characteristics and their influences on engagement," in *Proceedings of the 26th International Conference on Multimodal Interaction*, 2024, pp. 224–233.
- [185] M. J. Seikavandi, A. Modica, A. Obara, S. A. Shaffi, F. B. Narcizo, T. Ignatenko, T. Vucurevich, K. Haddad, D. Barratt, D. Overholt *et al.*, "Groupaffect-4: A multimodal dataset of four-person collaborative interaction," *arXiv preprint arXiv:2605.19765*, 2026.
- [186] Ropedia, "Xperience-10m: A large-scale egocentric multimodal dataset with structured 3d/4d annotations," Hugging Face, 2026, dataset.
- [187] J. Tao, F. Liu, M. Zhang, and H. Jia, "Design of speech corpus for mandarin text to speech," in *Proceedings of the Blizzard Challenge 2008 Workshop*, 2008.
- [188] C. Cui, Y. Ren, J. Liu, F. Chen, R. Huang, M. Lei, and Z. Zhao, "EMOVIE: A Mandarin Emotion Speech Dataset with a Simple Emotional Text-to-Speech Model," in *Interspeech 2021*, 2021, pp. 2766–2770.
- [189] J. Fan, W. Xu, O. Sokolsky, I. Lee, and F. Kong, "Safevla-bench: A benchmark for the success-safety gap in vision-language-action models," *arXiv preprint arXiv:2606.00773*, 2026.
- [190] H. Mai, B. Zhu, and T. Do, "Beyond task success: Behavioral and representational diagnostics for wam and vla," *arXiv preprint arXiv:2606.01095*, 2026.
- [191] G. Ferroni, N. Turpault, J. Azcarreta, F. Tuveri, R. Serizel, Ç. Bilen, and S. Krstulović, "Improving sound event detection metrics: insights from dcse 2020," in *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2021, pp. 631–635.
- [192] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
- [193] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.
- [194] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [195] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2, 2001, pp. 749–752.
- [196] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.
- [197] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," in *2020 twelfth international conference on quality of multimedia experience (QoMEX)*, 2020, pp. 1–6.

- [198] S. Kim, D. Le, W. Zheng, T. Singh, A. Arora, X. Zhai, C. Fuegen, O. Kalinli, and M. Seltzer, "Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric," in *Interspeech 2022*, 2022, pp. 3978–3982.
- [199] D. Thennal, J. James, D. P. Gopinath *et al.*, "Advocating character error rate for multilingual asr evaluation," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 4926–4935.
- [200] B. Phukon, X. Zheng, and M. Hasegawa-Johnson, "Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches," in *Interspeech 2025*, 2025, pp. 5708–5712.
- [201] S. Dixit, S. Deshmukh, and B. Raj, "Mace: Leveraging audio for evaluating audio captioning systems," in *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2025, pp. 1–5.
- [202] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva *et al.*, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.
- [203] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "Paradise: A framework for evaluating spoken dialogue agents," in *ACL*, 1997.
- [204] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak, "Survey on evaluation methods for dialogue systems," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 755–810, 2021.
- [205] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.
- [206] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification," in *The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 312–319.
- [207] T. Wang, P. Zheng, S. Li, and L. Wang, "Multimodal human-robot interaction for human-centric smart manufacturing: A survey," *Advanced Intelligent Systems*, vol. 6, no. 3, p. 2300359, 2024.
- [208] Z. Huang, Y. Shen, J. Li, M. Fey, and C. Brecher, "A survey on ai-driven digital twins in industry 4.0: Smart manufacturing and advanced robotics," *Sensors*, vol. 21, no. 19, p. 6340, 2021.
- [209] S. Li, P. Zheng, S. Liu, Z. Wang, X. V. Wang, L. Zheng, and L. Wang, "Proactive human-robot collaboration: Mutual-cognitive, predictable, and self-organising perspectives," *Robotics and Computer-Integrated Manufacturing*, vol. 81, p. 102510, 2023.
- [210] R. B. Rosenberg-Kima, Y. Koren, and G. Gordon, "Robot-supported collaborative learning (rscl): Social robots as teaching assistants for higher education small group facilitation," *Frontiers in Robotics and AI*, vol. 6, p. 148, 2020.
- [211] A. K. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: Pepper: The first machine of its kind," *IEEE Robotics & Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.
- [212] O. Engwall, J. Lopes, and A. Åhlund, "Robot interaction styles for conversation practice in second language learning," *International Journal of Social Robotics*, vol. 13, no. 2, pp. 251–276, 2021.
- [213] C. Zhou and F. Hou, "Can ai empower l2 education? exploring its influence on the behavioural, cognitive and emotional engagement of efl teachers and language learners," *European Journal of Education*, vol. 59, no. 4, p. e12750, 2024.
- [214] D. Abrahamson, M. J. Nathan, C. Williams-Pierce, C. Walkington, E. R. Ottmar, H. Soto, and M. W. Alibali, "The future of embodied design for mathematics teaching and learning," in *Frontiers in education*, vol. 5, 2020, p. 147.
- [215] D. Yang, E.-S. Oh, and Y. Wang, "Hybrid physical education teaching and curriculum design based on a voice interactive artificial intelligence educational robot," *Sustainability*, vol. 12, no. 19, p. 8000, 2020.
- [216] Amazon, "Amazon astro, household robot for home monitoring, with alexa," 2026.
- [217] E. Broadbent, K. Loveys, G. Ilan, G. Chen, M. Chilukuri, S. Boardman, P. Doraiswamy, and D. Skuler, "Elliq, an ai-driven social robot to alleviate loneliness: progress and lessons learned," *The Journal of Aging Research & Lifestyle*, vol. 13, pp. 22–28, 2024.
- [218] S. Coşar, M. Fernandez-Carmona, R. Agrigoroaie, J. Pages, F. Ferland, F. Zhao, S. Yue, N. Bellotto, and A. Tapus, "Enrichme: Perception and interaction of an assistive robot for the elderly at home," *International Journal of Social Robotics*, vol. 12, no. 3, pp. 779–805, 2020.
- [219] H. Cuayáhuil and G. Jang, "A dataset of robot-patient and doctor-patient medical dialogues for spoken language processing tasks," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 15 537–15 541.
- [220] H. W. Park, C. Breazeal, S. Alghowinem, A. K. Ostrowski, J. Ferguson, X. Zhang, and D. W. Lee, "Jibo community social robot research platform@ scale," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 1346–1348.
- [221] B. Friedman, P. H. Kahn Jr, and J. Hagman, "Hardware companions? what online aibo discussion forums reveal about the human-robotic relationship," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003, pp. 273–280.
- [222] B. Zhu, P. Schmitt, P. Meister, L. Gensler, M. Khalil, E. Poggi, J. Hechtl, C. Braunroth, K. Wurm, G. Narayanan *et al.*, "A factory-floor deployment case study of vla pipelines for industrial packaging task: Workflow, failures, and lessons," *arXiv preprint arXiv:2605.27461*, 2026.
- [223] H. Yamamoto, C. J. Mayer, C. Raithe, T. Buchner, C. Werner, Y. Hirata, M. Eckstein, and K. Mombaur, "Perception of social robots as communication partners in healthcare for older adults," *arXiv preprint arXiv:2605.21053*, 2026.
- [224] V. Pitardi, J. Wirtz, S. Paluch, and W. H. Kunz, "Service robots, agency and embarrassing service encounters," *Journal of service management*, vol. 33, no. 2, pp. 389–414, 2022.
- [225] L. Pham, T. Raza-Ullah, and H. Shaker, "Customer-robot interaction beyond the first encounter: a reciprocal framework and quadrant model for value co-creation," *International Journal of Consumer Studies*, vol. 50, no. 3, p. e70208, 2026.
- [226] J. Wang, L. Zheng, J. Wu, and Y. Mao, "Audio-visual world models: Towards multisensory imagination in sight and sound," *arXiv preprint arXiv:2512.00883*, 2025.