UbiHearo: Bringing Scenario-Aware Sound Guidance to DHH Users with Mobile Agents

FENGMIN WU, Northwestern Polytechnical University, China
SICONG LIU*, Northwestern Polytechnical University, China
ZIMU ZHOU, City University of Hong Kong, China
CHENREN XU, Peking University, China
WANBING ZHAO, Northwestern Polytechnical University, China
PEIWEN SUN, Northwestern Polytechnical University, China
ZHIWEN YU*, Harbin Engineering University, China

For Deaf and Hard-of-Hearing (DHH) users, effective sound awareness goes beyond detecting *what* sound occurs; it requires understanding *what it means*, *how urgent it is*, and *what to do*. However, existing mobile assistants largely reduce acoustic scenes to isolated labels or generic alerts, failing to provide contextual guidance when the same sound implies different risks across scenarios. We present UbiHearo, a mobile hearing assistant agent that closes the sensing–reasoning–action loop for *scenario-aware* and *personalized* DHH assistance. UbiHearo adopts a *neuro-symbolic* design that transforms continuous acoustic and physical context streams into structured intermediate representations (IRs) for controllable reasoning and guidance. It introduces three key mechanisms. *First*, a perception-thresholded sensing mechanism enables energy-efficient always-on awareness by activating processing only for perceptually meaningful events while preserving safety-critical cues through deterministic overrides. *Second*, a staged neuro-symbolic reasoning framework performs promptless scenario understanding by conditioning an on-device audio-language model on structured IRs instead of user-authored context descriptions. *Third*, a bounded human-in-the-loop personalization mechanism adapts user-specific guidance through lightweight on-device decision-boundary updates while preserving population-aligned reasoning. We evaluate UbiHearo on smartphones and in-vehicle systems across diverse scenarios. It achieves favorable accuracy–latency–energy tradeoffs and personalized guidance.

CCS Concepts: • **Human centered computing**; • **Ubiquitous and mobile computing**; • **Computing methodologies**; • **Machine learning**;

ACM Reference Format:
Fengmin Wu, Sicong Liu, Zimu Zhou, Chenren Xu, Wanbing Zhao, Peiwen Sun, and Zhiwen Yu. 2026. UbiHearo: Bringing Scenario-Aware Sound Guidance to DHH Users with Mobile Agents. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3, Article 170 (September 2026), 27 pages. https://doi.org/10.1145/3832010

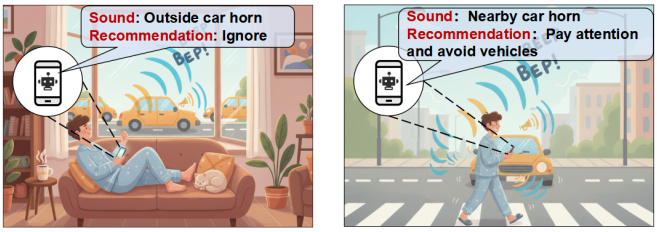
1 INTRODUCTION

Deaf and Hard-of-Hearing (DHH) users face significant challenges because many *safety-critical* and *socially meaningful* sounds are inaccessible to them. Missing a door knock, alarm, or shouted warning can jeopardize safety, while missing speech cues in shared spaces undermines social interaction. While hearing aids and cochlear

*Corresponding author: Sicong Liu, Zhiwen Yu.

Authors' Contact Information: Fengmin Wu, Northwestern Polytechnical University, China; Sicong Liu, Northwestern Polytechnical University, China; Zimu Zhou, City University of Hong Kong, China; Chenren Xu, Peking University, China; Wanbing Zhao, Northwestern Polytechnical University, China; Peiwen Sun, Northwestern Polytechnical University, China; Zhiwen Yu, Harbin Engineering University, China.

ACM 2474-9567/2026/9-ART170
https://doi.org/10.1145/3832010



(a) Indoor case (b) Outdoor case, street crossing

Fig. 1. Motivation cases, which illustrate scenario-dependent sound awareness for a DHH user, *Lucas*.

implants help restore auditory perception, they do not reliably provide **situational sound awareness** in dynamic mobile environments that vary across users, time, and locations. As a ubiquitous platform, smartphones provide built-in microphones and computing capabilities to continuously sense in-situ environmental sounds and deliver timely alerts, complementing hearing aids and serving as a fallback when they are unavailable. This has driven the development of mobile sound-awareness apps [8, 12, 18, 23, 29].

Despite advances, existing solutions still fall short of meeting DHH users' real needs in three critical areas, highlighted by our survey of 100 DHH users aged 15-60 (Sec. 2):

(1) **Sensing**. Many systems achieve high average recognition accuracy but still miss exactly the events DHH users care about most: *brief*, *low-SNR*, and *overlapping events* that occur *unpredictably* in daily life, *e.g.,* a shouted warning under traffic noise, a name call during a crowded commute, or an alarm mixed with TV/audio. This is exacerbated by energy-saving duty-cycled sensing, which can *skip* short transient cues.
(2) **Reasoning**. Most existing apps [8, 23] map mixed physical scenes into isolated sound labels, *e.g., car horn*. For DHH users, however, *the label is not the answer*: the same sound can indicate very different meaning, urgency, and required attention depending on *where they are*, *what they are doing*, and *how close/fast the source is*. For instance, a horn heard indoors may be irrelevant, while a horn during street crossing demands immediate action (Fig. 1). Without scenario-aware grounding, label-only alerts often lead to *false urgency* (unnecessary interruptions) or *missed urgency* (delayed reaction).
(3) **Guidance (Action)**. Label-only alerts shift interpretation burden back to DHH users, who must infer appropriate actions under time pressure and limited attention. Our survey shows strong demand for actionable situational guidance (*e.g.,* "check the door," "pause and look for traffic"), yet such assistance remains largely unexplored in existing mobile sound awareness systems.

These observations motivate a mobile hearing assistant that operates as a **sensing-reasoning-action loop**. It should continuously sense acoustic events under mobile energy budgets, interpret their situational meanings in *user-specific physical contexts*, and provide timely guidance aligned with individual needs.

However, achieving this goal faces three fundamental challenges:

• **Challenge #1:** Always-on awareness requires a difficult trade-off between recall and energy. Continuous sensing provides high recall for critical events but incurs substantial energy overhead [8], while aggressive duty cycling reduces energy consumption at the risk of missing brief and safety-critical sounds [28]. The challenge is determining *when and what to sense* to maintain continuous awareness under mobile energy constraints.

• **Challenge #2:** Sound awareness must answer not only *what sound is detected*, but also *what does it mean here*, and *what should I do now*, because physical context determines urgency and appropriate actions. Although LLMs (*e.g.,* GPT [25], Qwen [6]) provide strong reasoning capabilities, their performance depends heavily on *informative prompts* describing such context. However, DHH users in mobile scenarios often cannot provide fine-grained prompts due to time pressure, limited attention and environmental awareness. The challenge is enabling reliable scenario grounding and efficient reasoning without requiring user-authored prompts.

• **Challenge #3.** Scenario-aware reasoning captures population-level needs, but effective assistance is user-specific. Users may differ in preferred sound priorities, interruption tolerance, and such preferences may evolve over time. However, on-device personalization suffers from sparse and context-skewed feedback. Naive updates may overfit individual corrections, drift from population- and scenario-aligned knowledge, and destabilize safety-critical guidance. Existing DHH personalization approaches [8, 12, 18] mainly focus on sound-type customization and do not support scenario-conditioned preference adaptation.

To address these, we present UbiHearo, a mobile hearing assistant agent that provides *scenario-aware* and *personalized* sound awareness and guidance for DHH users. It includes three key mechanisms.

• **First.** We observe that human perception provides a natural sensing boundary: always-on awareness does not require always-on processing, as only perceptually meaningful events matter. Based on this insight, UbiHearo introduces *perception-driven sensing*, leveraging perceptual thresholds in spectrogram energy space to *filter* irrelevant audio via short-term thresholded listening, *extract* compact acoustic IRs via perceptual frequency decomposition and lightweight GMM modeling, and *activate* context sensing only upon meaningful events.

• **Second.** We observe that scenario-aware assistance requires hierarchical reasoning: grounding the user's physical context *before* interpreting sound meaning. Moreover, LLMs require informative prompts *mainly because* contextual states are not explicitly available during inference. Thus, UbiHearo introduces a *neuro-symbolic staged reasoning* mechanism. It first adopts a 1.5B-parameter Qwen2.5-Instruct model to infer physical context, then transforms heterogeneous context cues and acoustic tags into symbolic scenario states (*i.e.* IRs), which serve as *structured prompts* for an on-device 1.8B-parameter Qwen2-Audio model to generate scenario-aware guidance.

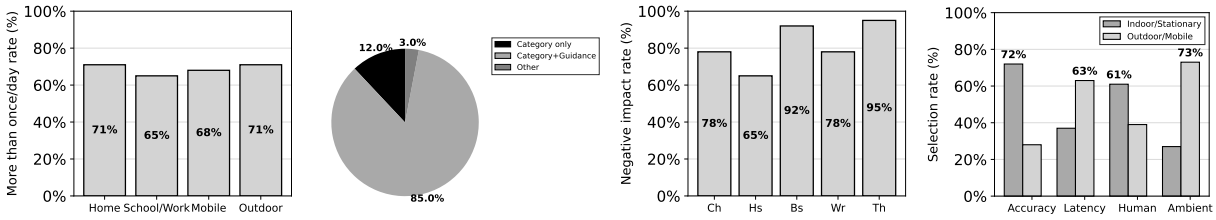
• **Third.** We note that personalization should adapt *how to respond* rather than *what to understand*, since acoustic semantics and physical context reasoning are shared while guidance preferences are user-specific. Harnessing this, UbiHearo updates a lightweight output-layer adapter from per-instance feedback, and leverages regularization to constrain updates as bounded deviations from population-aligned behavior, enabling private adaptation while reducing overfitting, preference drift, and online oscillation.

We prototype UbiHearo on smartphones and in-vehicle systems to deliver end-to-end sensing, scenario-aware notifications, and actionable guidance. Across diverse datasets and real-world tests spanning varied acoustic conditions and physical contexts, UbiHearo consistently outperforms five state-of-the-art baselines. It reduces miss rate by up to 10.3% while consuming 391× less energy, enabling up to 12.5 h of continuous operation under high-frequency trigger setting. For reasoning, UbiHearo improves scenario-aware user preference alignment by up to 61% via stage-wise task decomposition and scenario-conditioned preference training, while reducing model parameters by 53.8%. UbiHearo runs near real time on-device (*e.g.,* 0.9 s on Google Pixel 9), whereas audio-MLLM baselines (*e.g.,* Qwen-Audio2 [4], Audio-Flamingo [10], MiDashengLM [14]) remain too resource-intensive for on-device deployment. In summary, we make the following contributions:

- We identify and formalize a key requirement for DHH hearing assistance: moving beyond sound labeling to scenario-aware understanding and actionable physical guidance, grounded in a DHH user study.
- We design UbiHearo, a neuro-symbolic mobile assistant that unifies sensing, reasoning, and action through perceptual-thresholded duty-cycled sensing, structured intermediate representations, route-conditioned on-device reasoning, and human-in-the-loop personalization.
- We test UbiHearo on diverse mobile scenarios, showing that it improves scenario-aware reasoning, user-aligned guidance, and accuracy-latency-energy trade-offs, enabling practical near-real-time DHH assistance.

2 BACKGROUND AND MOTIVATION

To ground our design in real user needs, we conducted a questionnaire (Appendix A.1) study with 100 DHH participants (aged 15~60), combining multiple-choice and open-ended questions (Fig. 2). Permission was obtained



(a) Missing rates (b) Requiring guidance rates (c) Impact rates (d) Preference

Fig. 2. Survey results. (a) Missing rates of acoustic events, (b) Expectations for guidance, (c) Impact of universal guidance, and (d) Scenario-specific preference.

from the participants. The results reveal that effective mobile hearing assistance must go beyond sound labeling, and instead provide *scenario-aware interpretation*, *situational guidance*, and *instant personalized trade-offs*.

Formally, we define **Scenario** as a semantically grounded, user-centric *state* that links *i)* the user's location and activity, *ii)* the spatial relationship between the sound source and the user, and *iii)* the surrounding acoustic context (*e.g.,* foreground/background mixing of speech and ambient sounds). Scenario grounding is essential for translating detected sounds into appropriate user guidance, while dynamically controlling the system trade-off among accuracy, urgency, and resource efficiency under different contexts.

2.1 Survey Findings

2.1.1 Frequently Missed Speech and Ambient Sound Events. The survey shows that 91% of participants own a smartphone and are willing to use a mobile app for sound awareness, suggesting high acceptability. Yet reliable sound awareness remains difficult: 71%, 65%, 68%, and 71% report missing acoustic events more than once per day at home, at school/work, while mobile, and outdoors, respectively (Fig. 2a). Participants attributed these misses to short-duration cues, mixed speech and ambient sounds, noisy environments, and overlapping sources. These findings highlight the need for always-on mobile sensing that reliably captures context-critical events while maintaining energy efficiency.

2.1.2 Expectations Beyond Label Alerts. Participants need more than sound labels. As shown in Fig. 2b, 85% expect *situational guidance* after receiving alerts; among them, 79% (67/85) find labels alone insufficient to judge event urgency or determine appropriate actions, and 81% (69/85) believe situational guidance would improve safety. Consistent with prior psychological findings [2], participants prefer guidance such as "check the door" or "stay cautious near traffic" over isolated label (*e.g.,* "door knock"). Such guidance reduces cognitive burden when DHH users must make rapid decisions under limited sound awareness and attention.

2.1.3 Scenario Awareness Determines Appropriate Guidance. We further asked participants to compare *scenario-aware guidance* with *scenario-agnostic universal guidance* for five representative acoustic events (Tab. 1). Because the same sound may imply different urgency and actions across scenarios, scenario-agnostic universal guidance can be ineffective or misleading. As shown in Fig. 2c, participants reported such issues for car horns (78%), human shouting (65%), broadcast sounds (92%), running water (78%), and thunder (95%). These highlight the *essential role* of scenario-aware reasoning and guidance.

2.1.4 Scenario-varying User Preferences. DHH users' preferences dynamically vary across scenarios (Fig. 2d). When indoors and stationary, 72% prioritize higher accuracy, whereas outdoors and mobile, 63% prioritize faster responsiveness. Their event priorities also shift with physical context: 73% emphasize safety-critical ambient sounds outdoors (*e.g.,* traffic), while 61% emphasize human speech and socially meaningful cues indoors. These show that user preferences depend on context and individual needs, making fixed configurations suboptimal.

Table 1. Examples of guidance strategies in the Questionnaire for common acoustic events across various scenarios

Acoustic event	Scenario	Scenario-aware guidance	Assumed-universal guidance
Car horn(Ch)	Indoor	Ignore	Ignore
	On the road	Attention to vehicle	
Human shouting(Hs)	Playground	Ignore	Ignore
	Outdoor	Attention to source	
Broadcast sound(Bs)	At school/square	Ignore	Ignore
	Railway station/Airport	Attention to departure time	
Water running(Wr)	Home	Turn off the tap	Turn off the tap
	Wild	Attention to water source	
Thunder(Th)	Indoor	Ignore	Ignore
	Outdoor	Attention to lighting	

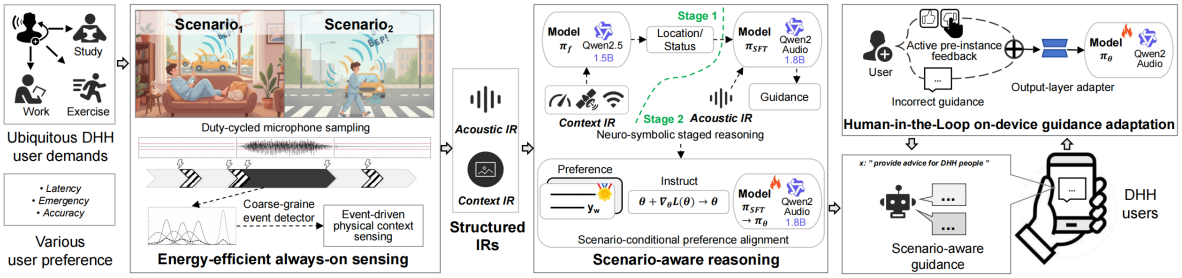


Fig. 3. Workflow of UbiHearo.

2.2 Motivating Example

Consider *Lucas*, a DHH user who relies on a mobile assistant for sound awareness. At home, *Lucas* is stationary and prefers accurate recognition, while many non-human sounds are low priority. When a car horn is heard outside the building (Fig. 1a), the assistant should recognize the event, understand that it poses limited urgency in this indoor context, and avoid unnecessary interruptions. Later, when *Lucas* walks outdoors, responsiveness becomes critical because traffic sounds may indicate immediate safety risks. If the same car horn occurs near a street crossing (Fig. 1b), the assistant should detect it promptly and provide actionable guidance (*e.g.*, checking surroundings) while continuing to monitor subsequent cues. This example reveals a fundamental requirement: **scenario context determines both the meaning of an acoustic event and the appropriate system response**. Therefore, practical mobile hearing assistance must adapt its sensing strategy and guidance according to changing scenarios.

2.3 Design Goals

Guided by the above findings, a mobile hearing assistant should satisfy three design goals:

- **Energy-efficient always-on sensing.** Survey results (Sec. 2.1.1) show that missed speech and ambient sound events pose safety risks. The system should support continuous acoustic monitoring under strict battery constraints while preserving sensitivity to user-critical events.
- **Scenario-aware sensing-reasoning-action.** Findings from Sec. 2.1.2 and Sec. 2.1.3 show that sound meaning and appropriate responses depend on scenario context. Instead of producing isolated labels, the system should integrate acoustic signals with physical context (*e.g.*, location, motion, and activity) to infer scenarios, reason about urgency, and provide actionable guidance.
- **On-device human-in-the-loop personalization.** DHH users' preferences are highly individual, scenario-dependent, and evolve over time (Sec. 2.1.4), making static models insufficient. The system should support on-device, human-in-the-loop personalization, enabling incremental adaptation via user feedback and lightweight fine-tuning without cloud retraining or privacy risks.

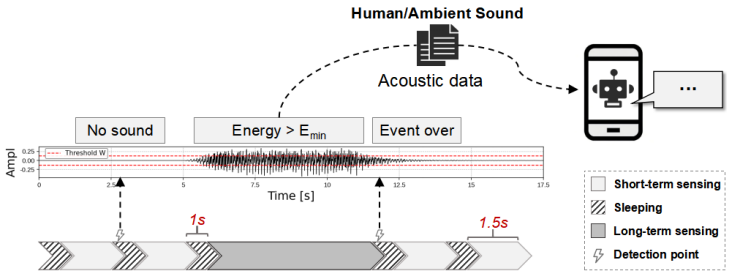


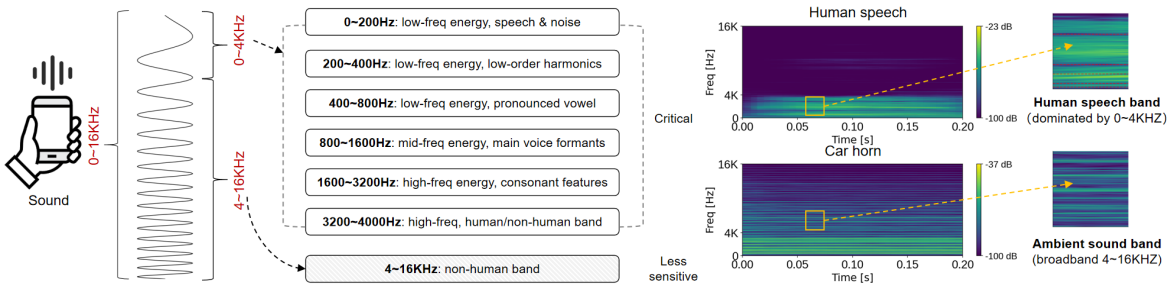
Fig. 4. Illustration of the duty-cycled microphone sampling mechanism.

3 SYSTEM DESIGN

3.1 Overview

We design UbiHearo as a mobile hearing assistant that transforms continuous acoustic and contextual signals into scenario-aware notifications and actionable guidance through a **sensing–reasoning–action pipeline**. UbiHearo adopts a **neuro-symbolic design** [24], where lightweight sensing modules extract reliable evidence from the environment, intermediate representations (IRs) provide symbolic states for scenario grounding, and an aligned audio-language model generates user-facing guidance. This design moves beyond sound classification and black-box multimodal reasoning by enabling controllable scenario-level interpretation.

- **Sensing.** To enable energy-efficient always-on awareness, UbiHearo employs a perception-thresholded sensing mechanism inspired by the *Weber–Fechner law* [36]. The key insight is that always-on awareness does not require always-on processing: UbiHearo performs lightweight *short-term listening* to suppress silence and irrelevant noise, and activates *long-term listening* only when perceptually meaningful events emerge; otherwise, it enters *sleep* to reduce idle energy. Rather than making final decisions, this module extracts reliable acoustic evidence from continuous audio streams through perceptually motivated frequency analysis and a lightweight GMM front end. Verified safety-critical events bypass reasoning and directly trigger deterministic alerts (Sec. 3.2).
- **Structured Intermediate Representation (IR)**. Raw sensing outputs alone are insufficient for scenario understanding because acoustic meaning depends on physical context. UbiHearo therefore converts sensing outputs into two IRs: an *acoustic IR* describing event evidence and safety relevance, and a *context IR* describing physical states such as mobility and indoor/outdoor status. These IRs provide a symbolic interface between sensing and reasoning, enabling automatic scenario grounding without requiring DHH users to provide context descriptions under time pressure and limited attention.
- **Reasoning**. UbiHearo reformulates DHH scenario reasoning from end-to-end black-box inference into a *neuro-symbolic staged reasoning process*. Instead of directly mapping raw multimodal inputs to responses, UbiHearo first infers physical scenarios from lightweight sensor cues and then reasons about acoustic events conditioned on structured IRs. The resulting IR-guided prompts enable an on-device audio-language model to generate scenario-aware notifications and guidance while reducing multimodal ambiguity and supporting scenario-conditioned preference alignment (Sec. 3.3).
- **Action (guidance).** UbiHearo maps inferred scenarios into concise guidance (*e.g., ignore*, *investigate*, or *act cautiously*). Beyond static rules or cloud-based retraining, UbiHearo enables *human-in-the-loop on-device adaptation*: user feedback on incorrect interpretations or guidance is locally converted into user-specific supervision signals. Instead of updating the entire model, UbiHearo adapts only a lightweight output-layer adapter, localizing personalization at the final decision boundary while preserving general acoustic and scenario reasoning (Sec. 3.4).



(a) Seven non-uniform frequency sub-bands (b) The spectrogram of human talk and car horn

Fig. 5. The basis for non-uniform frequency band division.

3.2 Energy-Efficient Always-On Sensing

3.2.1 Duty-Cycled Microphone Sampling. DHH users require continuous sound awareness, yet most environmental audio contains no meaningful information. Existing approaches face a recall–energy dilemma: continuous microphone processing provides reliable detection but consumes excessive power [8], whereas fixed or coarse-grained adaptive sampling may miss short-lived alarms, sirens, or verbal calls [12, 18]. UbiHearo resolves this by grounding sensing decisions in *perceptual detectability* rather than fixed sampling schedules. Our key observation is that sound perception has an intrinsic energy boundary: acoustic events below the just-perceptible energy threshold ($E_{\min} \approx 10^{-6}$ J) are fundamentally inaudible to humans [35]. Therefore, sensing should preserve events that accumulate sufficient perceptual energy instead of continuously processing all audio. For a sound with duration T and perceived bandwidth B, its accumulated energy can be approximated as:

$$E \propto \sum_{n} \int_{B} |x(n,f)|^2 df \cdot \Delta t, \tag{1}$$

where $|x(n,f)|^2$ denotes the energy density and Δt represents the temporal integration window. This formulation provides a perceptual basis for mapping human sound perception into sensing decisions.

Based on this principle, UbiHearo adopts a duty-cycled listening mechanism. Because DHH-relevant events are temporally sparse, UbiHearo first performs *short-term listening* for inexpensive event discovery. During this stage, it suppresses silence and background noise and produces *coarse* acoustic cues. When a non-silent event is detected, UbiHearo switches to *long-term listening* to capture richer acoustic information; otherwise, it enters *sleep* mode to reduce idle energy. Importantly, the sensing layer does not make final semantic decisions. Instead, it filters irrelevant audio and converts continuous streams into compact cues for downstream scenario reasoning. UbiHearo maps perceptual constraints into sensing parameters as follows:

- *Sampling rate.* UbiHearo uses 32 kHz sampling to preserve spectral information within the main human-audible frequency range and support subsequent event verification.
- *Frame length.* UbiHearo adopts 20 ms frames as the energy integration window, balancing stable short-time energy estimation and low detection latency.
- *Duty-cycled schedule.* UbiHearo samples audio in 10-s cycles at 0, 2.5, 5, and 7.5 s. Intervals that fail to accumulate sufficient energy toward $E_{\min}$ are skipped, reducing unnecessary sensing with minimal perceptual loss.
- *Short-term listening duration.* UbiHearo dynamically adjusts short-term listening duration according to battery level: [0.02–1.60 s] when battery > 50% and [0.02–1.00 s] otherwise. These bounds remain above perceptual integration requirements while limiting sensing overhead.

3.2.2 Coarse-grained Acoustic Event Detector. After perceptual-thresholded sensing identifies potentially meaningful audio, UbiHearo extracts compact acoustic evidence rather than performing expensive semantic classification. The goal is not to determine final sound meaning, but to provide reliable and interpretable acoustic cues for downstream scenario reasoning. Therefore, UbiHearo employs a lightweight GMM detector instead of neural sound classifiers (*e.g.,* YAMNet [13], Silero [32]). The detector exploits an acoustic asymmetry between speech and ambient sounds: speech exhibits structured energy patterns in low- and mid-frequency bands, whereas ambient sounds have more diverse spectral distributions. Accordingly, UbiHearo divides each frame into seven non-uniform frequency sub-bands (Fig. 5a), assigning finer resolution to lower bands where speech-related cues are more perceptually salient under the *Weber-Fechner law* [36]. As shown in Fig. 5b, speech energy is mainly concentrated in 0–4 kHz, while ambient sounds spread more broadly across 0–16 kHz. For each frame t, UbiHearo computes the energy of sub-band i as:

$$E_{t,i} = \int_{B_i} |X_t(f)|^2 df, \quad x_{t,i} = \log(E_{t,i} + \epsilon), \tag{2}$$

where B_i denotes the i-th sub-band, $X_t(f)$ represents the short-time spectrum, and ϵ ensures numerical stability. UbiHearo models each sub-band log-energy distribution using a GMM:

$$p_i(x_{t,i}) = \sum_{j=1}^{K_i} w_{i,j} \mathcal{N}\left(x_{t,i} \mid \mu_{i,j}, \sigma_{i,j}^2\right), \tag{3}$$

where $w_{i,j}$, $\mu_{i,j}$, and $\sigma_{i,j}^2$ denote the mixture weight, mean, and variance of the j-th component. Since the GMM operates in log-energy space, the expected physical energy is estimated as:

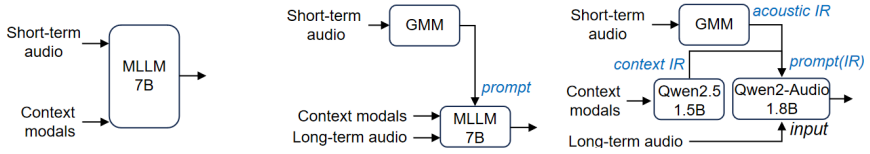
$$\bar{E}_i = \mathbb{E}_{p_i}\left[\exp(x)\right] = \sum_{j=1}^{K_i} w_{i,j} \exp\left(\mu_{i,j} + \frac{1}{2}\sigma_{i,j}^2\right). \tag{4}$$

During online detection, UbiHearo first checks whether observed energy exceeds the perceptual threshold τ_E derived from $E_{\min}$:

$$E_t = \sum_{i=1}^{M} E_{t,i} = \sum_{i=1}^{M} \int_{B_i} |X_t(f)|^2 df. \tag{5}$$

where M is the number of sub-bands. For detected events, UbiHearo uses GMM likelihoods $\{p_i(x_{t,i})\}_{i=1}^{M}$ to estimate speech consistency. High-energy frames with low speech consistency are conservatively classified as ambient sounds, reducing false speech detections. The detector outputs an *acoustic IR* containing event type, confidence score, and safety relevance, which serves as the interface to scenario reasoning (Sec. 3.3). For safety-critical sounds, UbiHearo adds a deterministic verification layer using band-wise energy ratios and temporal consistency (Appendix A.2). Verified alarms and sirens bypass LLM reasoning and directly trigger alerts; uncertain events receive conservative notifications.

3.2.3 Event-Driven Physical Context Sensing. Scenario reasoning also requires physical context, but continuously sensing all modalities wastes energy because environmental context evolves more *slowly* than acoustic events. Thus, UbiHearo adopts event-driven context acquisition: it activates physical sensing only after detecting perceptually meaningful acoustic events. Upon an acoustic trigger, UbiHearo extracts three lightweight context cues: *step frequency* for mobility estimation, *Wi-Fi density and SSID semantics* for indoor context inference, and *satellite visibility with SNR* for outdoor confirmation. These robust physical indicators [34] are mapped into structured context labels, forming a *context IR*. Together, the acoustic IR and context IR provide automatically generated structured inputs for downstream scenario reasoning.



(a) Scheme 1 (b) Scheme 2 (c) UbiHearo (Two-stage)

Fig. 6. Comparison of three reasoning schemes. (a) A single MLLM directly processes all modalities. (b) A 7B MLLM reasons over GMM-derived prompts. (c) UbiHearo uses two lightweight models, combining contextual states and GMM outputs as structured prompts.

3.3 Scenario-Aware Reasoning

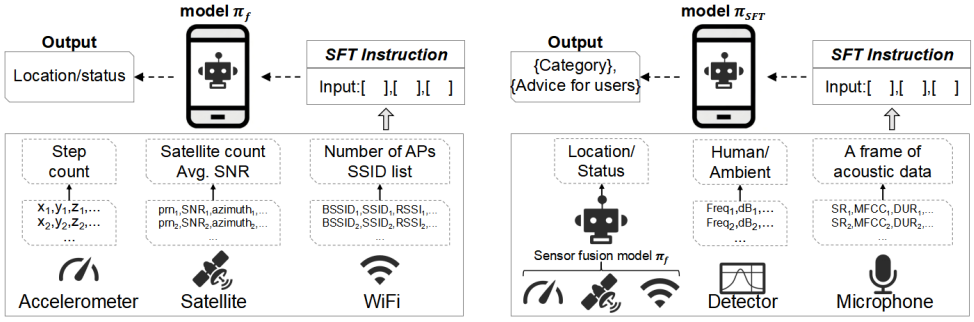
Effective DHH assistance requires grounding acoustic evidence into user-specific scenarios and generating appropriate responses. However, directly applying generic audio-language models to mobile DHH assistance is insufficient. *First*, large multimodal models are difficult to deploy under mobile latency and energy constraints. *Second*, feeding raw audio and heterogeneous sensor streams directly into a generic model entangles perception and reasoning, increasing inference cost and hallucination risks. *Third*, models trained on general hearing-population data lack awareness of DHH users' scenario-dependent priorities, such as when to ignore, investigate, or immediately react to an event. UbiHearo addresses these through an IR-guided two-stage reasoning architecture. The key *insight* is that scenario reasoning should be explicitly grounded before acoustic interpretation.

3.3.1 Neuro-symbolic Staged On-Device Reasoning. UbiHearo follows two *observations* to design its reasoning pipeline. *First*, physical context inference and acoustic interpretation represent different reasoning dimensions. Physical states, such as mobility and indoor/outdoor status, are relatively low-dimensional and can be inferred from lightweight sensors, whereas acoustic meaning requires richer semantic reasoning. Separating these two processes avoids forcing a single model to jointly learn heterogeneous modalities and reduces unnecessary reasoning complexity. *Second*, compact models become effective when their inputs are structured and aligned with user preferences. Prior work shows that preference alignment can enable smaller models to outperform larger generic models in human evaluations [26]. Therefore, instead of directly prompting a large model with raw observations, UbiHearo provides structured IRs that encode scenario-relevant information.

Based on these observations, UbiHearo adopts a two-stage IR-guided on-device reasoning pipeline (Fig. 6c), where sensing-generated IRs automatically replace manual user prompts.

- *Stage 1: Physical context inference.* UbiHearo first infers structured physical states from lightweight sensor cues. A compact modality-fusion model based on Qwen2.5-1.5B-Instruct processes accelerometer signals, Wi-Fi features, and satellite visibility to infer context labels, such as *indoor/outdoor* and *stationary/mobile*. These labels form an explicit scenario scaffold for downstream Stage 2.
- *Stage 2: Scenario reasoning and guidance generation.* Given the inferred context IR and acoustic IR, UbiHearo constructs structured prompts for an on-device audio-language model. The model is derived from Qwen2-Audio-7B-Instruct and compressed to 1.8B parameters through hierarchical quantization [17] and low-rank decomposition [15]. It jointly reasons over acoustic evidence, physical context, and acoustic tags to generate scenario-aware notifications and actionable guidance.

Why Structured IRs Help. From a *neuro-symbolic* perspective, IRs serve as explicit scenario states rather than hand-crafted rules. They bridge low-level perception and high-level reasoning by transforming continuous multimodal observations into structured, low-entropy inputs. Moreover, LLM reasoning heavily depends on informative prompts, which are difficult for DHH users to provide during mobile interaction due to limited environmental awareness, attention, and time pressure. UbiHearo therefore automatically constructs *structured prompts* through IRs, organizing user situations into scenario routes such as *indoor/stationary*, *indoor/mobile*,



(a) SFT process of modal-fusion model π_f (b) SFT process of scenario reasoning model π_{SFT}

Fig. 7. Illustration of the stage-wise supervised fine-tuning.

outdoor/stationary, and *outdoor/mobile*. These routes guide reasoning and align responses with scenario-specific priorities. Thus, UbiHearo enables reliable scenario interpretation without requiring users to manually describe their context, while allowing a general audio-language model to perform scenario-specialized reasoning.

3.3.2 Training Strategy. Scenario-aware reasoning requires both stable scenario grounding and alignment with DHH users' preferences. However, directly optimizing an end-to-end model may blur the boundary between context inference, acoustic interpretation, and guidance generation, while generic preference optimization cannot capture scenario-dependent needs. Therefore, UbiHearo adopts a *two-step* training strategy: *stage-wise supervised fine-tuning* (SFT) establishes functional interfaces between reasoning stages, and *scenario-conditioned direct preference optimization* (DPO) aligns final guidance with DHH-specific preferences.

Stage-wise Supervised Fine-Tuning. SFT aims to provide each model with stable and interpretable capabilities before preference alignment. We train the context inference model and scenario reasoning model separately using role-specific triplets $(instruction, input, output)$ (Fig. 7). The context model π_f takes sampled sensor cues as input and outputs structured physical states, such as *indoor/outdoor* and *stationary/mobile*. The scenario reasoning model π_{SFT} takes inferred context labels, GMM-derived acoustic tags, and raw audio as input, and generates scenario-aware notifications and physical guidance. The SFT objective is:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log P\left(y_t \mid x, y_{<t}\right), \tag{6}$$

where x denotes the input signals and instruction, y_t represents the target token, and T denotes the output length. After SFT, UbiHearo obtains a grounded context model π_f and an initial guidance model π_{SFT}. However, SFT mainly captures population-level reasoning patterns and cannot fully model how DHH users trade off urgency, interruption, and responsiveness across scenarios.

Scenario-Conditioned Preference Alignment. DHH assistance requires scenario-dependent preferences: the same sound may require different responses depending on whether the user is indoors or outdoors, stationary or mobile. For example, a distant car horn outside a building may be safely ignored, while the same sound near a street crossing requires immediate attention. Therefore, UbiHearo performs *preference alignment* after scenario routing. The structured prompt maps each instance into a scenario slot, including: *indoor/stationary*, *indoor/mobile*, *outdoor/stationary*, and *outdoor/mobile*. This converts mixed-context preference learning into route-conditioned alignment. Within each scenario slot, we model two key preference dimensions: *sound priority* and *accuracy–latency trade-off*. Preference pairs are constructed from the DHH user study and scenario-specific guidance principles. For each input x, the preferred response y_w is concise, context-aware, and action-oriented. The rejected response y_l violates one or more principles, such as ignoring physical context, misjudging urgency,

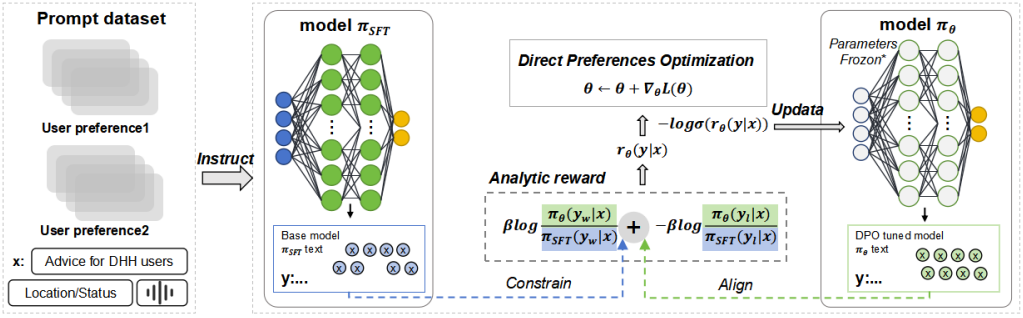


Fig. 8. Illustration of the scenario-conditioned preference alignment mechanism.

omitting recommended actions, or prioritizing inappropriate sounds. Ambiguous cases are excluded from DPO training, while safety-critical events follow the deterministic override path described in Sec. 3.2.2.

We adopt Direct Preference Optimization (DPO) instead of RLHF [3], as it avoids explicit reward modeling and rollout-based optimization. Starting from π_{SFT}, DPO optimizes the preference margin between y_w and y_l:

$$\mathcal{L}_{\text{DPO}}\left(\pi_\theta; \pi_{\text{ref}}\right) = -\mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{SFT}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{SFT}}(y_l \mid x)}\right)\right] \tag{7}$$

where π_θ is initialized from π_{SFT}, $\pi_{\text{ref}} = \pi_{\text{SFT}}$, and β controls deviation from the reference model.

Example Preference Cases. We illustrate scenario-conditioned DPO using non-safety-critical events (Table 2). Safety-critical sounds such as sirens and alarms are excluded because they bypass LLM reasoning through deterministic override. Across these cases, preferred responses are selected because they are concise, scenario-aware, and action-oriented, whereas rejected responses are context-insensitive, overly aggressive, or behaviorally misleading.

Table 2. Examples of scenario-conditioned preference pairs.

Scenario	**Preferred response y_w**	**Rejected response y_l**
Indoor, stationary; distant car horn	A distant car horn is detected outside. Since you are indoors and stationary, no immediate action is needed.	A car horn is detected. Stop what you are doing and immediately check for nearby traffic.
Outdoor, mobile; nearby car horn	A nearby car horn is detected. Stay alert, slow down, and check surrounding vehicles before moving.	A car horn is detected. It is common background noise and can be ignored.
Indoor, mobile; nearby speech	Nearby speech is detected while you are walking. Slow down and check whether someone is calling you.	Speech is detected indoors. It is likely irrelevant background conversation, so keep walking.
Outdoor, stationary; repeated honks	Repeated vehicle honks are detected nearby. Stay attentive and check traffic conditions before moving.	Repeated honks are normal urban noise. You can safely disregard them completely.

3.4 Human-in-the-Loop On-Device Guidance Adaptation

Although Sec. 3.3 aligns UbiHearo with population-level DHH preferences, assistive guidance remains user-specific. The same acoustic event may require different guidance across users, routines, and environments.

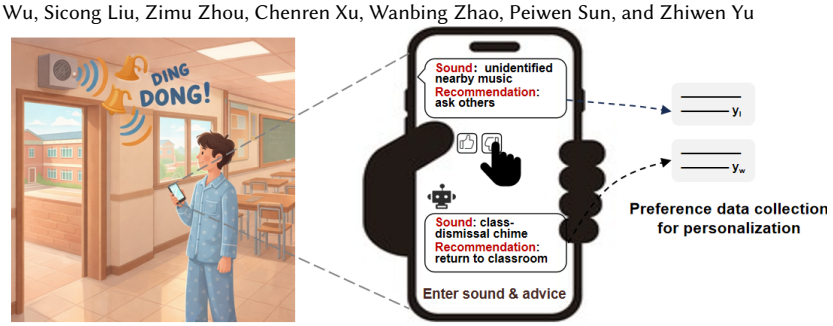


Fig. 9. Illustration of the customized configurations.

Existing personalization approaches often rely on confidence-triggered feedback or cloud-side retraining, which may miss confident-but-wrong decisions and introduce privacy and deployment overhead. UbiHearo treats personalization as *controlled decision correction* rather than full-model re-optimization. It learns lightweight user-specific adjustments while preserving population-aligned reasoning through constrained on-device adaptation.

3.4.1 Active Per-Instance Feedback. Confidence alone cannot identify user-specific mismatches, as models may produce confident but inappropriate guidance. Therefore, UbiHearo enables per-instance feedback after each response. Users can correct both event interpretation and recommended actions, with each correction stored locally as a supervision pair for targeted adaptation. This transforms personalization from global retraining into selective error correction.

3.4.2 Efficient Adaptation with Output-Layer Adapter. On-device personalization must adapt to individual preferences without destabilizing general reasoning. Naive fine-tuning may overfit sparse feedback or overwrite previous corrections. Thus, UbiHearo freezes the backbone models and updates only a lightweight output-layer adapter using LoRA-style low-rank updates. This localizes adaptation at the decision boundary while preserving acoustic and scenario reasoning. Formally, UbiHearo adjusts the output logits as:

$$\pi_{\text{final}}(y \mid x) = \pi_\theta(y \mid x) + \alpha\, \Delta_U(x), \tag{8}$$

where π_θ denotes the frozen population model, Δ_U represents the user-specific adapter, and α controls adaptation strength. The adapter is optimized with a stability-regularized objective:

$$\mathcal{L}_{\text{user}} = \mathcal{L}_{\text{DPO}}(\pi_{\text{final}}, \mathcal{D}_u) + \lambda \cdot \text{KL}(\pi_{\text{final}} \parallel \pi_\theta) + \eta \cdot \mathcal{R}_{\text{stab}}, \tag{9}$$

where $\mathcal{D}_u$ is the local feedback set. The KL constraint preserves population-aligned behavior, while $\mathcal{R}_{\text{stab}}$ prevents repeated rewriting of similar decisions. Together, they make personalization a bounded correction rather than behavioral drift. For efficiency, UbiHearo updates only FP16 adapter parameters. A stabilization window after each update further reduces oscillation and self-overwriting during continuous adaptation.

3.4.3 Default and Customized Operation. UbiHearo supports two modes. UbiHearo supports both default and customized operation. In *default* mode, pretrained models provide scenario-aware guidance without user intervention. In *customized* mode, user corrections trigger incremental local adapter updates. As shown in Fig. 9, a high school student, *Lucas*, initially receives generic guidance when broadcast music is interpreted as an undefined ambient sound. After correcting it as a class-bell chime and specifying the desired action (*"return to the classroom"*), UbiHearo learns this preference and provides more appropriate guidance for similar situations.

4 EVALUATION

This section presents the experimental settings and system performance of UbiHearo.

Table 3. Quantized model footprint for reasoning modules.

Model	Params	Quantization	Raw weight memory
π_f fusion model	1.5B	2-bit	~0.38GB
π_θ audio-language model	1.8B	2-bit	~0.45GB
UbiHearo reasoning stack	3.3B	2-bit	~0.83GB
Qwen-Audio2	7B	2-bit	~1.75GB
Audio-Flamingo	8B	2-bit	~2.00GB
MiDashengLM	8B	2-bit	~2.00GB

4.1 Experiment Setups

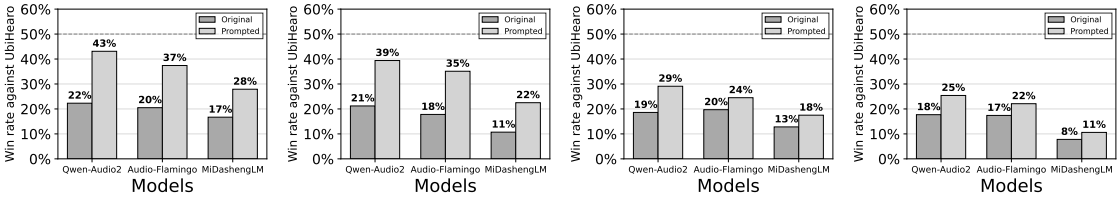
Implementation. UbiHearo is implemented as a two-stage pipeline. Offline training (SFT, DPO) and ablation studies are conducted on a server with an NVIDIA RTX 3090 GPU and 256GB RAM. The online system is deployed as an Android application on a Google Pixel 8 for continuous background sensing and multi-modal reasoning. Offline training and ablation studies use FP16, while on-device deployment uses 2-bit weight-only quantization for the reasoning modules. Table 3 summarizes the quantized footprint of UbiHearo and the reasoning baselines.

On-device deployment. We deploy both UbiHearo modules on a Google Pixel 8 with 2-bit weight-only quantization. The 1.5B fusion model π_f and the 1.8B audio-language model π_θ are co-located on device but executed sequentially in an event-driven pipeline: after a non-silent acoustic trigger, π_f infers context labels, and π_θ generates guidance from these labels and the long-listened audio. Their quantized footprints are approximately 0.38 GB and 0.45 GB, respectively, making the combined stack feasible on Pixel 8-class hardware. For fairness, we apply the same 2-bit quantization to all LLM baselines, but report full Android deployment only for UbiHearo because deploying 7B–8B audio-LLM baselines requires additional model-specific kernel and memory optimizations.

Tasks and Models. We use a **self-collected physical context sensory dataset** from a Google Pixel 8 (1K+ clips over 7 days), including 18K accelerometer readings, 3.5K Wi-Fi AP counts, 2.8K SSID entries, and 4,200 satellite counts with SNR values, to supervise π_f. We also use AudioSet [9] for acoustic training and large-scale evaluation. We evaluate UbiHearo on two tasks: coarse detection (T_1) and multi-modal context-aware guidance (T_2), measuring both task performance and on-device energy cost.

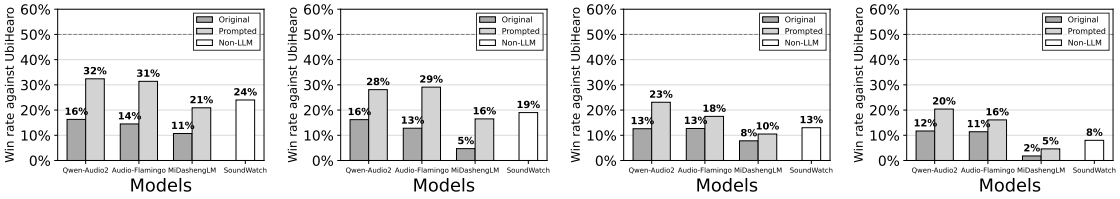
Baselines. Because UbiHearo covers both sensing and guidance generation, we compare it with representative baselines from two families.

- **Lightweight DNN-based sound detection methods** benchmark T_1 under different accuracy, latency, and energy trade-offs.
  - *Silero* [32]: an edge-oriented neural voice activity detector with low parameter cost.
  - *YAMNet* [13]: a CNN-based audio event model from Google, pre-trained on large-scale audio data and used as a stronger but heavier detection baseline.
- **LLM and non-LLM reasoning methods** benchmark T_2 for scenario-aware guidance.
  - *SoundWatch* [20]: a non-LLM VGG-lite sound classification baseline. Since it outputs labels rather than guidance, we convert its predictions into fixed rule-based notifications for pairwise evaluation.
  - *Qwen-Audio2* (7B) [4]: an audio-language model combining Whisper-large-v3 with Qwen-7B for speech, audio analysis, and mixed-scenario reasoning.
  - *Audio-Flamingo* (8B) [10]: an open-source audio-language framework supporting long-audio understanding across speech, music, and environmental sounds.
  - *MiDashengLM* (8B) [14]: an audio-language model designed for low-resource acoustic scenarios and efficient zero-/few-shot audio-text generation.



(a) Subset D_1 (b) Subset D_2 (c) Subset D_3 (d) Subset D_4

Fig. 10. UbiHearo's (π_θ) pairwise preference win rate on diverse scenarios.



(a) Subset D_1 (b) Subset D_2 (c) Subset D_3 (d) Subset D_4

Fig. 11. UbiHearo's ($\pi_f + \pi_\theta$) pairwise preference win rate on diverse scenarios.

Together, these baselines cover the key axes of *efficiency vs. generalization* and *task-specific modeling vs. general-purpose reasoning*, allowing us to evaluate how UbiHearo balances on-device efficiency with scenario-aware, preference-aligned assistance.

Metrics. We evaluate UbiHearo along three dimensions: (1) **Detection quality**, using Precision-Recall (PR) curves across SNR levels; (2) **System efficiency**, using on-device latency and energy consumption; and (3) **Guidance quality**, using pairwise preference win rates. The win rate is obtained through blind pairwise comparisons, with Qwen3-Omni (30B) [5] as the primary automated judge for large-scale guidance evaluation. Human participants are involved separately in the 21-day field study and the 100-user quick test, where they assess perceived usefulness, understandability, and actionability rather than serving as judges for the large-scale pairwise benchmark.

4.2 Performance of UbiHearo's Reasoning Module

As described in Sec. 3.3, UbiHearo replaces monolithic audio-MLLMs with a *two-stage* reasoning pipeline. A compact modal-fusion model π_f (1.5B) first infers structured user context from on-device sensors, and a lightweight audio-language model π_θ (1.8B) then performs scenario reasoning over long-listened audio. The GMM acoustic tags and inferred context labels are used as structured prompts to π_θ, with raw audio as the primary input. We evaluate scenario-aware reasoning using *pairwise preference win rate* (Sec. 4.1) on four AudioSet-derived subsets (2,000 clips): D_1 indoor/stationary isolated sounds, D_2 outdoor/mobile isolated sounds, D_3 indoor/stationary speech+ambient overlap, and D_4 outdoor/mobile speech+ambient overlap. We compare (i) π_θ alone and (ii) the full pipeline ($\pi_f + \pi_\theta$) against Qwen-Audio2 (7B) [4], Audio-Flamingo (8B) [10], and MiDashengLM (8B) [14], each in both original and prompted forms.

4.2.1 Performance of Scenario Reasoning Model π_θ. We first evaluate π_θ under identical inputs and prompts (*GMM outputs + context labels*). We conduct blind pairwise preference evaluation using Qwen3-Omni as the automated judge, where the judge selects the more useful and actionable guidance for each clip under randomized output order. As shown in Fig. 10, UbiHearo consistently exceeds 50% win rate against all baselines. On D_1 and

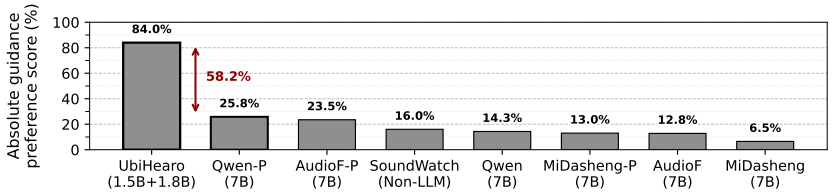


Fig. 12. Absolute guidance preference score converted from the pairwise win rates in Fig. 11. For each baseline b on subset D_i, its score is its win rate against UbiHearo, *i.e.* $S_{b,i} = W_{b,i}$. For UbiHearo, we report the mean complementary win rate against all baselines, $S_{\text{UbiHearo},i} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} (100 - W_{b,i})$. The final score is averaged over D_1–D_4. "P" denotes the prompted version. Higher is better.

D_2, it wins 57% and 61% against the strongest baseline, prompted Qwen-Audio2. The gain is larger in overlapped and dynamic settings: on D_4, UbiHearo reaches up to 92% win rate, while MiDashengLM drops to 8%. These results show that SFT+DPO enables π_θ to produce more scenario-aware and preference-aligned guidance under challenging acoustic conditions.

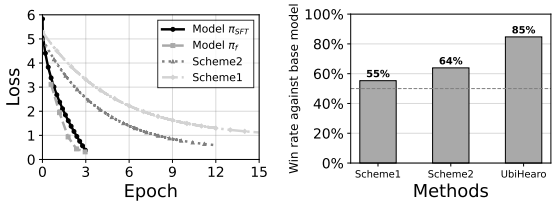
4.2.2 Performance of UbiHearo's *Two-stage Pipeline (π_f+π_θ).* We then evaluate the full pipeline on the same four subsets. All LLM-based methods use the same long-listened audio and prompts containing *GMM outputs + raw on-device sensor readings*. Unlike the baselines, UbiHearo first converts raw sensors into structured context labels through π_f, testing whether explicit context grounding improves guidance beyond model scale. We also include a SoundWatch-style non-LLM baseline [20]. Since SoundWatch outputs labels rather than guidance, we convert labels into fixed notifications: *"Detected [sound label]. Please check the source if needed."* We further add SoundWatch+Rule, which maps selected safety-related labels to simple actions, such as alarm/siren/horn → check surroundings, door knock/doorbell → check the door, water running → check the source, and otherwise → be aware. As shown in Fig. 11, UbiHearo outperforms all baselines on every subset. On D_1, it wins 68% against the strongest LLM baseline, which reaches 32%. On D_2, UbiHearo achieves 71%–95% win rates under more dynamic conditions. Against the SoundWatch-style baseline, UbiHearo wins 76%, 81%, 87%, and 92% on D_1–D_4, respectively, with larger gains in overlapped and outdoor-mobile scenarios where label-only or rule-based notifications struggle to infer urgency. For clearer interpretation, we convert the win rates in Fig. 11 into an absolute guidance preference score (Fig. 12). UbiHearo achieves 84.0%, substantially higher than the strongest baseline, prompted Qwen-Audio2, at 25.8%.

Summary. Overall, UbiHearo improves DHH-oriented scenario understanding through structured context grounding and preference-aligned sensing–reasoning–action, enabling a compact two-stage model to outperform larger audio-MLLMs and non-LLM baselines.

4.3 Deep Dive into UbiHearo's Two-stage Reasoning Scheme

As described in Sec. 3.3.1, UbiHearo uses stage-wise SFT to separate context understanding from guidance generation. In Stage 1, the context model π_f maps on-device sensor signals to structured context labels. In Stage 2, the guidance model π_{SFT} takes a structured prompt, composed of coarse acoustic tags and context labels, together with long-listened audio clips to generate scenario-aware guidance. By contrast, Scheme 1 directly feeds *sensor signals + short-listened audio* into a single-stage model, while Scheme 2 uses *acoustic tags + sensor signals + long-listened audio* without explicit context structuring.

4.3.1 Training Efficiency. We compare UbiHearo with Scheme 1 and Scheme 2 by fine-tuning the same base model (Qwen2-Audio-7B-Instruct) under identical SFT settings. As shown in Fig. 13a, both π_f and π_{SFT} converge



(a) The loss trajectories of SFT (b) Diverse schemes

Fig. 13. Performance of UbiHearo's two-stage reasoning scheme.

within *3 epochs*, faster than Scheme 1/2, and reach lower losses (0.18 and 0.36) than Scheme 1 (0.59) and Scheme 2 (1.11). This shows that decomposing reasoning into context structuring and audio-grounded guidance reduces learning difficulty and improves training efficiency.

4.3.2 Downstream Reasoning Quality. We further evaluate guidance quality using pairwise preference win rate across Scheme 1, Scheme 2, UbiHearo, and the base model (Fig. 13b). Despite using fewer parameters (1.5B+1.8B *vs.* 7B), UbiHearo achieves an 84.7% win rate over the base model, showing that two-stage reasoning with automatically generated structured prompts improves preference-aligned guidance beyond single-stage prompting.

Summary. Overall, UbiHearo's two-stage reasoning scheme improves both training convergence and scenario-aware guidance quality.

4.4 Performance of UbiHearo's Acoustic Event Detector

We evaluate UbiHearo's acoustic event detector on a *Google Pixel 8* against two mobile baselines, YAMNet [13] and Silero [32], focusing on noise robustness and always-on efficiency. We test three SNR levels (15/20/25 dB) using 14,800 audio frames from 50 AudioSet clips covering *Human Speech* and *Ambient Sound*.

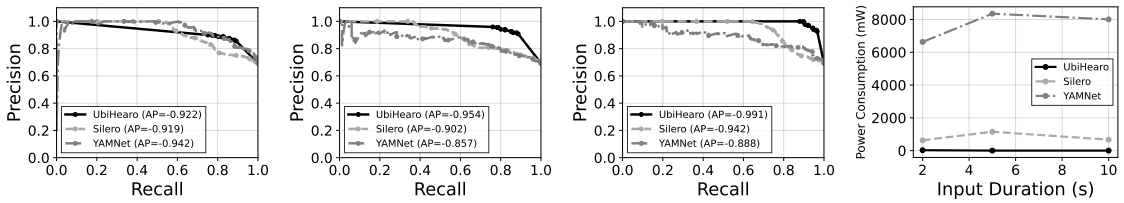
4.4.1 Detection Accuracy under Different SNRs. Fig. 14a-Fig. 14c report precision-recall curves and average precision (AP). UbiHearo achieves AP scores of 0.922/0.954/0.991 across 15/20/25 dB, matching or outperforming both baselines at 20–25 dB and remaining slightly below YAMNet only at the near-threshold 15 dB condition. Its runtime memory footprint is only 158 KB, ∼369× smaller than YAMNet (57 MB), supporting continuous sensing under tight mobile resource budgets.

4.4.2 Energy-Accuracy Trade-off. We further measure on-device power at 20 dB SNR and compare UbiHearo with YAMNet and Silero. As shown in Fig. 14d, UbiHearo provides the best accuracy-energy trade-off: it consumes only 20 mW, ∼391× lower than YAMNet (7,836 mW), while maintaining competitive detection precision (AP ≈ 0.956 averaged across SNRs). Energy is measured on a Google Pixel 8 using *Android Battery Historian* and *dumpsys batterystats* under the same always-on sensing condition. Before each run, we reset battery statistics and closed unrelated background applications. These results show that UbiHearo enables accurate and energy-efficient coarse acoustic detection for always-on mobile sensing.

4.5 End-to-end System Performance

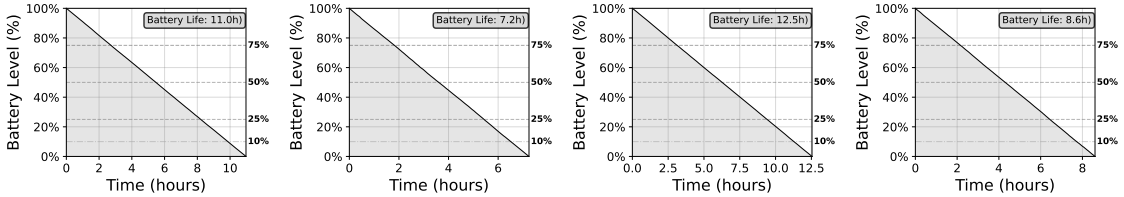
We evaluate UbiHearo's end-to-end overhead on Google Pixel 8 (4575 mAh, 8 GB RAM) and Pixel 9 (4700 mAh, 12 GB RAM) to examine its feasibility for *always-on* mobile sensing.

4.5.1 Power Consumption and Battery Life. We measure end-to-end power consumption by repeatedly triggering UbiHearo with locally stored AudioSet clips, avoiding network effects. Specifically, we use ten 3-s *Human Speech* clips and ten 3-s *Ambient Sound* clips, keep the Pixel 8 screen off unless guidance is displayed, and randomly



(a) PR curve under 15dB (b) PR curve under 20dB (c) PR curve under 25dB (d) Consumption vs. Duration

Fig. 14. Performance of acoustic event detector.



(a) Once/duty-cycling at pixel 8 (b) Twice/duty-cycling at pixel 8 (c) Once/duty-cycling at pixel 9 (d) Twice/duty-cycling at pixel 9

Fig. 15. System overhead of UbiHearo at Google pixel 8 and pixel 9 smartphones.

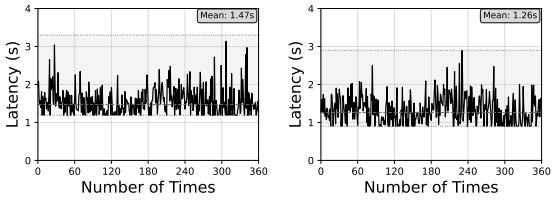
pair each trigger with one of four scenario-specific sensor instances for context inference by π_f. We test two conservative trigger frequencies, once and twice per duty cycle, both more frequent than typical daily acoustic events. As shown in Fig. 15a and Fig. 15c, UbiHearo achieves 11.0 h battery life on Pixel 8 and 12.5 h on Pixel 9 under once-per-duty-cycle triggering. When triggered twice per duty cycle, battery life decreases to 7.2 h and 8.6 h, respectively (Fig. 15b). These results indicate that UbiHearo can support practical always-on use.

4.5.2 System Latency. We report end-to-end processing latency under the once-per-duty-cycle setting, measured between the fourth and fifth hour of the battery test to capture steady-state performance. As shown in Fig. 16a and Fig. 16b, UbiHearo achieves near-real-time average latency: 1.47 s on Pixel 8 and 1.26 s on Pixel 9. Existing audio-MLLM baselines, such as Qwen-Audio2, Audio-Flamingo, and MiDashengLM, are too large for fully on-device always-on deployment and therefore cannot provide comparable latency under the same setting. Latency also varies by scenario: it is lowest for *outdoor/mobile* cases, reaching 1.2 s on Pixel 8 and 0.9 s on Pixel 9, and highest for *indoor/stationary* cases, reaching 3.3 s and 2.20 s, respectively. This is because UbiHearo generates more concise guidance in outdoor/mobile scenarios and more detailed responses in indoor/stationary scenarios, as described in Sec. 3.3.2. While UbiHearo achieves near real-time latency on the tested Pixel devices, its latency may still vary with device capability, runtime workload, and response length. Future work will further optimize scheduling and output-length control towards real-time processing across broader deployment settings.

Summary. UbiHearo achieves energy-efficient always-on operation and near-real-time on-device performance on commodity smartphones, enabling timely notification in everyday use.

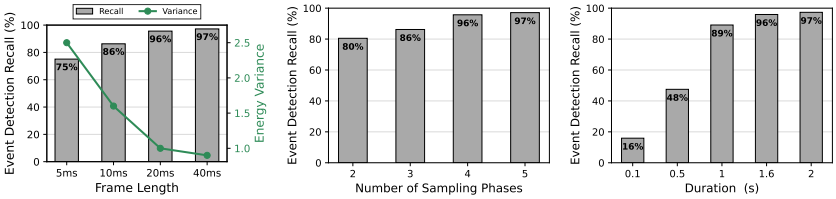
4.6 Micro-Benchmark

We conduct micro-benchmarks to quantify key design trade-offs in UbiHearo. We concatenate 10-s AudioSet clips into long streams and randomly insert 2–10 s silent gaps to emulate sparse always-on listening. The streams include speech, alarms, and short transient events, and we report *Event Detection Recall*, defined as the fraction of ground-truth events successfully detected.



(a) Latency at pixel 8 (b) Latency at pixel 9

Fig. 16. Latency of UbiHearo at Google pixel 8 and pixel 9.



(a) Frame length (b) Duty-cycling configurations (c) Listening duration

Fig. 17. Micro-benchmark of UbiHearo's configurations.

4.6.1 Frame Length. We evaluate four frame lengths (5–40 ms) to balance temporal granularity and stability (Fig. 17a). Short frames (5/10 ms) suffer from high energy variance and lower recall (75.1%/86.3%), while longer frames improve robustness. Although 40 ms slightly improves recall over 20 ms (97.2% *vs.* 95.6%), it doubles detection latency. We therefore adopt 20 ms as the accuracy-latency trade-off.

4.6.2 Duty-cycled Schedule. We evaluate 2–5 sampling phases per 10 s interval (Fig. 17b). Sparse schedules with 2/3 phases miss transient events and reduce recall (80.5%/86.2%). Four phases increase recall to 95.7%, while five phases bring only marginal gain (97.1%) with higher overhead. Thus, we use four phases by default.

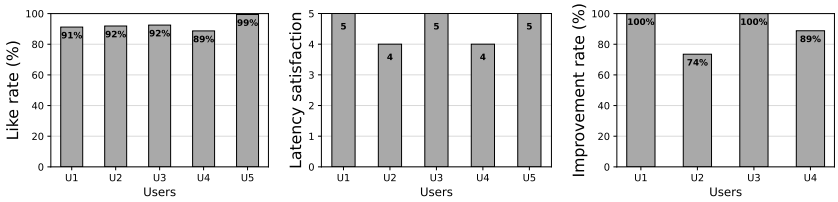
4.6.3 Listening Duration. We vary the listening duration from 0.1–2 s per trigger (Fig. 17c). A 1 s window already provides strong recall and is used in power-saving mode (battery $\leq$ 50%). A 1.6 s window improves recall to 95.9% and is used in standard mode. Thus, UbiHearo adopts an adaptive policy: 1 s for low-power operation and 1.6 s for higher-fidelity sensing.

Summary. These micro-benchmarks show that 20 ms frames, four duty-cycled phases, and adaptive listening duration provide a practical balance among recall, latency, and energy cost.

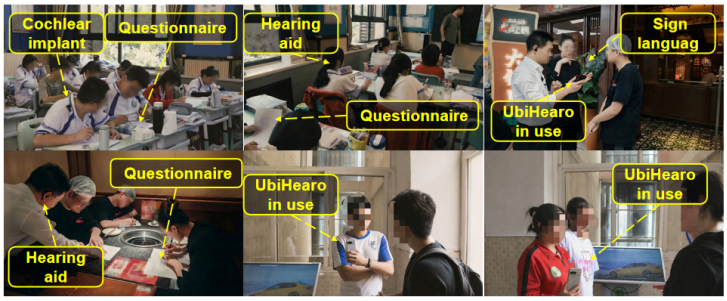
4.7 User Study

We conducted a two-tier user study to evaluate whether UbiHearo provides useful guidance in daily life, not only accurate predictions in controlled tests. We employed five DHH participants for a 21-day longitudinal field study and 100 additional participants for a breadth-side quick test.

21-Day Daily Usage. We employed five DHH participants: *U1* (age 19, congenitally deaf, student), *U2* (age 32, hard of hearing, employed), *U3* (age 27, hard of hearing, homemaker), *U4* (age 45, congenitally deaf, self-employed), and *U5* (age 55, hard of hearing, unemployed). We asked them to use UbiHearo in daily life for 21 days. During the study, UbiHearo was triggered *8,834*, *9,443*, *5,278*, *6,811*, and *3,514* times for *U1-U5*, respectively. To reduce burden, we asked participants to tap the on-screen "*dislike*" button (Fig. 9) only when the output did not match the current situation. As shown in Fig. 18a, UbiHearo achieved high user-rated appropriateness over 21 days:



(a) Guidance satisfaction (b) Latency satisfaction (c) Adaptation effect



(d) Examples of the questionnaire and user study process

Fig. 18. User study.

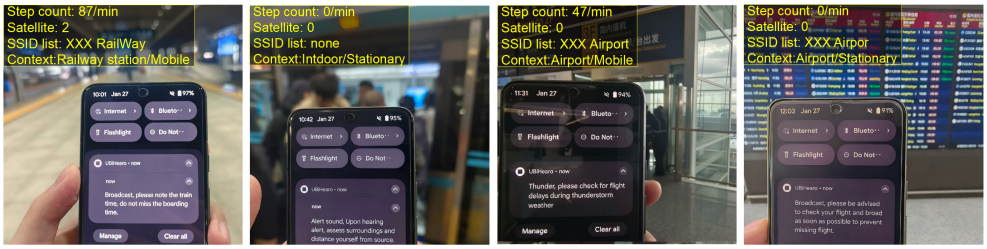
91.2% (8058/8834), 91.9% (8670/9443), 92.5% (4884/5278), 88.7% (6043/6811), and 99.4% (3492/3514) for *U1–U5*. We attribute the variation across users to differences in daily environments. As shown in Fig. 18b, participants rated UbiHearo's latency as 5, 4, 5, 4, and 5, with an average score of 4.6.

Human-in-the-loop On-device Adaptation. We asked four participants (*U1–U4*) to enable the *customized* mode and adapt UbiHearo to personal sound cues that generic datasets rarely cover, including a campus chime for class transitions (*U1*), TV audio at home (*U2*), microwave sounds (*U3*), and cash-register clicks at checkout (*U4*). As shown in Fig. 18c, adaptation improved recognition of these personalized events: *U1* 0/22→23/23, *U2* 0/30→25/34, *U3* 0/3→3/3, and *U4* 0/64→32/36. These results show that on-device human-in-the-loop adaptation helps UbiHearo learn user-specific sound meanings in real use.

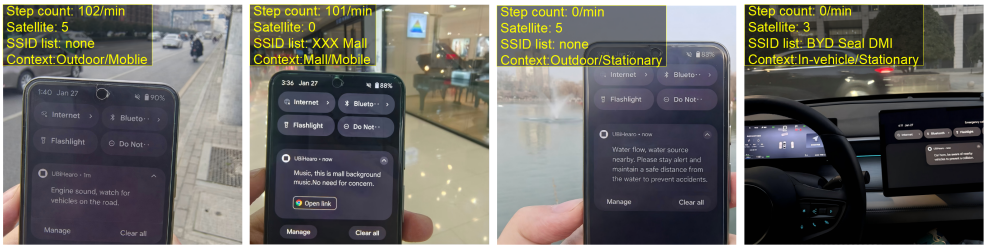
100-User Test. We employed 100 additional participants from campus and local communities for a lightweight quick test. We asked each participant to review four representative guidance clips, one for each scenario slot: indoor/stationary, indoor/mobile, outdoor/stationary, and outdoor/mobile. This yielded 400 clip-level judgments. For each clip, we asked participants to rate whether the scenario understanding was correct and whether the guidance was actionable. Overall, UbiHearo achieved 89.5% scenario-correct judgments (358/400) and 87.3% actionable judgments (349/400), showing that its guidance remained understandable and actionable beyond the five long-term participants. The 100-user quick test complements the 21-day field study by providing a broader controlled review of representative guidance clips. While the field study captures repeated real-world use, the quick test evaluates whether the guidance is understandable and actionable for a broader participant pool.

4.8 Case Study

To examine UbiHearo in real-world conditions, we deploy it on a Pixel 8 across everyday environments, including railway stations, subways, airports, roadside traffic, shopping malls, parks, and in-vehicle settings. These scenarios vary in mobility, background noise, interference, and salient sounds such as broadcasts, alarms, and horns. As shown in Fig. 19, we log representative on-device responses to assess whether the guidance is timely, scenario-aware, and actionable.



(a) Railway station (b) Subway (c) Airport, walking (d) Airport, stationary



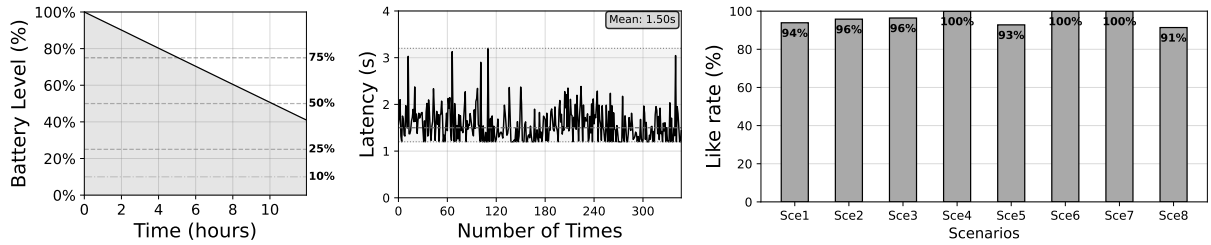
(e) Road (f) Mall (g) Park (h) In-vehicle

Fig. 19. In-the-wild case study on Pixel 8 across eight everyday scenarios, spanning diverse mobility, noise, and acoustic conditions (*e.g.,* interference, broadcasts, alarms, and horns) to assess timely, scenario-aware, actionable guidance.

In-the-moment examples. UbiHearo provides context-aware guidance under noisy and time-critical conditions. On a train platform, it maps boarding broadcasts to time-sensitive advice: *Broadcast, please note the train time and do not miss the boarding time* (Fig. 19a). In a crowded subway, a doors-close warning triggers safety guidance: *Alert sound, please assess your surroundings and keep a safe distance from the source* (Fig. 19b). At an airport, thunder and boarding reminders lead to flight-aware suggestions (Fig. 19c, Fig. 19d). Near dense traffic, engine noise triggers: *Engine sound, watch for vehicles on the road* (Fig. 19e). In benign continuous-sound settings, such as mall music and park fountains, UbiHearo correctly de-escalates notifications (Fig. 19f, Fig. 19g). Finally, when preparing to pull out of a parking spot and a nearby vehicle honks, UbiHearo issues a collision-aware warning (Fig. 19h).

Real-world system performance. We further quantify continuous-use practicality with Pixel 8 running only the OS and UbiHearo, with the screen off except when triggered. We log energy, latency, and output quality, and manually assess quality using the criteria in Sec. 3.3.2: sound-type priorities, accuracy-latency trade-offs, and high-priority sounds. Users mark appropriate outputs via an on-screen "*like*" button (Fig. 9). After ~12 hours of continuous testing, the device retained 41% battery (Fig. 20a), and across 347 triggers, UbiHearo achieved 1.5 s mean end-to-end latency (Fig. 20b).

The 347 instances cover eight location labels: *Sce1* Indoor (115), *Sce2* Outdoor (73), *Sce3* Railway station (28), *Sce4* Airport (34), *Sce5* Mall (14), *Sce6* Restaurant (20), *Sce7* Hospital (16), and *Sce8* In-vehicle (47). As shown in Fig. 20c, user-rated appropriateness remains high across scenarios: 93.9%, 95.8%, 96.4%, 100%, 92.8%, 100%, 100%, and 91.4%, respectively. Failures are explainable and concentrated in a few cases, such as weak satellite/WiFi signals limiting broadcast-to-arrival association, traffic-light *ding-ding* sounds misread as alerts, train whistles confused with wind, video-wall audio misclassified as speech, and door slams during vehicle entry/exit not reliably recognized. Speech-dominant environments, such as restaurants and hospitals, also bias detections toward speech, although most outputs still satisfy our criteria.



(a) Power consumption (b) System latency (c) Guidance quality (user-rated appropriateness)

Fig. 20. Real-world system performance.

5 RELATED WORK

5.1 Assistive Technologies and Human Augmentation for DHH Users

Mobile human augmentation aims to restore perceptual access through lightweight, real-time assistance. Prior DHH systems have explored AR/VR spatial visualization [16, 22], wearable haptics [7, 33], and IoT-based environmental indicators [19]. However, these solutions may suffer from hardware bulk, tactile fatigue, or limited portability. In contrast, UbiHearo uses ubiquitous smartphones to support a hardware-free sensing–reasoning–action loop, moving from basic sound awareness to scenario-aware guidance.

5.2 Transitioning Mobile Sound Awareness from Classification to Reasoning

Mobile sound awareness has evolved from handcrafted features and CNN classifiers [1, 29] to self-attention encoders [11] and large pretrained audio models [30]. Although these methods improve recognition, they mostly produce closed-set labels and can degrade in noisy, dynamic settings [18, 20]. Recent personalization methods further support on-device adaptation [8, 12], but still focus mainly on recognition. LLMs offer a shift from fixed labels to open-set reasoning [27]. UbiHearo builds on this shift by grounding acoustic tags with physical context, enabling scenario-aware guidance rather than ambiguous sound notifications.

5.3 Human Preference Alignment Techniques

Preference alignment is important for safety-related assistance, where outputs must reflect user priorities and acceptable actions. RLHF [21] is widely used but introduces reward modeling and online optimization costs that are unsuitable for lightweight mobile deployment. DPO [31] provides a more efficient alternative by directly optimizing preference likelihoods. UbiHearo adopts DPO to align generative guidance with DHH users' safety priorities and lifestyle preferences while keeping adaptation lightweight.

6 CONCLUSION

This paper addresses a key unmet need in mobile hearing assistance for Deaf and Hard-of-Hearing (DHH) users: moving beyond recognition-only sound labels toward *always-on, scenario-aware* understanding and *actionable guidance*. We present UbiHearo, a mobile sound guidance agent that closes the *sensing–reasoning–action* loop through energy-efficient duty-cycled sensing, structured acoustic/context representations, staged on-device reasoning, and lightweight human-in-the-loop personalization. By automatically generating structured prompts from acoustic tags and physical context, UbiHearo reduces the burden of user-crafted prompts and enables scenario-aware guidance in everyday environments. Across mobile devices and real-world tasks, UbiHearo improves guidance quality while maintaining practical energy and latency costs. Future work will incorporate richer low-power context signals and explore more efficient alignment and distillation methods to further reduce runtime cost while preserving user-aligned guidance.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (62522215, 62532009, 62472354, U25B2040, 62272010). Zimu Zhou's research is supported by CityU APRC grant (No. 9610633).

References

[1] Semih Agcaer, Anton Schlesinger, Falk-Martin Hoffmann, and Rainer Martin. 2015. Optimization of amplitude modulation features for low-resource acoustic scene classification. In *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2556–2560.

[2] Erin E Campbell and Elika Bergelson. 2022. Making sense of sensory language: Acquisition of sensory knowledge by individuals with congenital sensory impairments. *Neuropsychologia* 174 (2022), 108320.

[3] Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2025. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *Comput. Surveys* 58, 2 (2025), 1–37.

[4] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-Audio Technical Report. *arXiv preprint arXiv:2407.10759* (2024).

[5] Alibaba Cloud. 2024. Qwen3-Omni-30B-A3B-Instruct: A Universal Multi-Modal Large Language Model with Enhanced Audio-Visual Capabilities. https://www.modelscope.cn/models/Qwen/Qwen3-Omni-30B-A3B-Instruct. Accessed: 2026-01-15.

[6] Alibaba Cloud. 2026. Qwen. https://qianwen.com/. Accessed: 2026-01-30.

[7] Caluã de Lacerda Pataca, Saad Hassan, Lloyd May, Michelle M Olson, Toni D'aurio, Roshan L Peiris, and Matt Huenerfauth. 2025. Tactile Emotions: Multimodal Affective Captioning with Haptics Improves Narrative Engagement for d/Deaf and Hard-of-Hearing Viewers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[8] Hang Do, Quan Dang, Jeremy Zhengqi Huang, and Dhruv Jain. 2023. AdaptiveSound: An Interactive Feedback-Loop System to Improve Sound Recognition for Deaf and Hard of Hearing Users. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–12.

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

[10] Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128* (2025).

[11] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).

[12] Steven M Goodman, Emma J McDonnell, Jon E Froehlich, and Leah Findlater. 2025. SPECTRA: Personalizable Sound Recognition for Deaf and Hard of Hearing Users through Interactive Machine Learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[13] Google Research. 2019. YAMNet: A Pre-trained Model for Audio Event Classification. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet. Version 1.0; Accessed: 2026-01-09.

[14] Horizon Team, MiLM Plus. 2025. *MiDashengLM: Efficient Audio Understanding with General Audio Captions*. Technical Report. Xiaomi Inc. arXiv:2508.03983 https://arxiv.org/abs/2508.03983 Contributors: Heinrich Dinkel et al. (listed alphabetically in Appendix B).

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[16] Jeremy Zhengqi Huang, Jaylin Herskovitz, Liang-Yuan Wu, Cecily Morrison, and Dhruv Jain. 2025. Weaving Sound Information to Support Real-Time Sensemaking of Auditory Environments: Co-Designing with a DHH User. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[17] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2704–2713.

[18] Dhruv Jain, Khoa Huynh Anh Nguyen, Steven M. Goodman, Rachel Grossman-Kahn, Hung Ngo, Aditya Kusupati, Ruofei Du, Alex Olwal, Leah Findlater, and Jon E. Froehlich. 2022. Protosound: A personalized and scalable sound recognition system for deaf and hard-of-hearing users. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16.

[19] Dhruv Jain, Angela Lin, Rose Guttman, Marcus Amalachandran, Aileen Zeng, Leah Findlater, and Jon Froehlich. 2019. Exploring sound awareness in the home for people who are deaf or hard of hearing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[20] Dhruv Jain, Hung Ngo, Pratyush Patel, Steven Goodman, Leah Findlater, and Jon Froehlich. 2020. SoundWatch: Exploring smartwatch-based deep learning approaches to support sound awareness for deaf and hard of hearing users. In *Proceedings of the 22nd International*

ACM SIGACCESS Conference on Computers and Accessibility. 1–13.
[21] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback. (2024).
[22] Jaewook Lee, Davin Win Kyi, Leejun Kim, Jenny Peng, Gagyeom Lim, Jeremy Zhengqi Huang, Dhruv Jain, and Jon E Froehlich. 2025. SonoCraftAR: Towards Supporting Personalized Authoring of Sound-Reactive AR Interfaces by Deaf and Hard of Hearing Users. *arXiv preprint arXiv:2508.17597* (2025).
[23] Sicong Liu, Zimu Zhou, Junzhao Du, Longfei Shangguan, Jun Han, and Xin Wang. 2017. UbiEar: Bringing Location-independent Sound Awareness to the Hard-of-hearing People with Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2 (2017), 17–1.
[24] Minghui Ma, Bin Guo, Mengqi Chen, Jingqi Liu, Yasan Ding, Yan Liu, and Han Wang. 2026. Neuro-Sym Supporter: A Thoughtful Emotion Support Agent Integrating Neural and Symbolic Policy Learning. In *Proceedings of the ACM Web Conference 2026*. 3823–3834.
[25] OpenAI. 2026. ChatGPT. https://chatgpt.com/. Accessed: 2026-01-30.
[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
[27] Chunjong Park, Chulhong Min, Sourav Bhattacharya, and Fahim Kawsar. 2020. Augmenting conversational agents with ambient acoustic contexts. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–9.
[28] Rajesh Maharudra Patil and CM Patil. 2024. Unveiling the State-of-the-Art: A Comprehensive Survey on Voice Activity Detection Techniques. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*. IEEE, 1–5.
[29] Karol J Piczak. 2015. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*. IEEE, 1–6.
[30] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.
[31] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.
[32] Silero Team. 2021. Silero VAD: Pre-trained Enterprise-Grade Voice Activity Detector. https://github.com/snakers4/silero-vad. Version 3.1; Accessed: 2026-01-09.
[33] Yiwen Wang, Ziming Li, Pratheep Kumar Chelladurai, Wendy Dannels, Tae Oh, and Roshan L Peiris. 2023. Haptic-captioning: using audio-haptic interfaces to enhance speaker indication in real-time captions for deaf and hard-of-hearing viewers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.
[34] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*. 1–7.
[35] Lin Zhong and Niraj K Jha. 2005. Energy efficiency of handheld computer interfaces: limits, characterization and practice. In *Proceedings of the 3rd international conference on Mobile systems, applications, and services*. 247–260.
[36] Eberhard Zwicker and Hugo Fastl. 2013. *Psychoacoustics: Facts and models*. Vol. 22. Springer Science & Business Media.

A Appendix

In this appendix, we present the questionnaire used in the survey study described in Sec. 2, along with the algorithm for the coarse-grained acoustic event detector introduced in Sec. 3.2.

A.1 Questionnaire

(1) Which channels do you usually use to obtain sound information? (Single choice)
Cochlear implant:
□ Always □ Sometimes □ Occasionally □ Rarely □ Never
Hearing aid:
□ Always □ Sometimes □ Occasionally □ Rarely □ Never
Someone informs you via sign language:
□ Always □ Sometimes □ Occasionally □ Rarely □ Never
Ask friends or family:
□ Always □ Sometimes □ Occasionally □ Rarely □ Never
Use a phone app:
□ Always □ Sometimes □ Occasionally □ Rarely □ Never
Use a watch app:
□ Always □ Sometimes □ Occasionally □ Rarely □ Never

(2) How often do you miss sounds you are interested in? (Single choice)
At home:
□ More than once/day □ Once/day □ Once/week □ Once/month □ None
At work/school:
□ More than once/day □ Once/day □ Once/week □ Once/month □ None
When you are moving and looking at your phone:
□ More than once/day □ Once/day □ Once/week □ Once/month □ None
When you are moving but not looking at your phone:
□ More than once/day □ Once/day □ Once/week □ Once/month □ None

(3) Preference differences across scenarios (Single choice for each item)
Scenario A: Indoors and stationary
In this scenario, which factor do you prioritize the most for a sound-awareness assistant? (Single choice)
□ Recognition accuracy (whether the sound is correctly recognized)
□ Response latency (how quickly the prompt is delivered)
□ Sound category coverage (whether it can recognize the sounds I care about)
□ Other:
Scenario B: Outdoors and mobile
□ In this scenario, which factor do you prioritize the most for a sound-awareness assistant? (Single choice)
□ Recognition accuracy (whether the sound is correctly recognized)
□ Response latency (how quickly the prompt is delivered)
□ Sound category coverage (whether it can recognize the sounds I care about)
□ Other:
In addition, in these two scenarios, which type of sounds do you care about the most? (Single choice for each item)
Scenario A: Indoors and stationary (Single choice)

□ Human speech and socially meaningful sounds (*e.g.,* conversations, someone calling my name)
□ Safety-critical sounds (*e.g.,* alarms, gas leak alerts)
□ Home environment sounds (*e.g.,* , doorbell, knocking, appliance sounds)
□ Other:
Scenario B: Outdoors and mobile (Single choice)
□ Safety-critical ambient sounds (*e.g.,* traffic sounds, horns, approaching vehicles)
□ Human speech and socially meaningful sounds (*e.g.,* people calling, public announcements)
□ Natural environment sounds (*e.g.,* wind, rain, thunder)
□ Other:
Finally, do you think your preferences for the assistant (*e.g.,* accuracy vs. speed, sound categories) will change over time due to daily routines, occupation, or familiarity with the assistant? (Single choice)
□ Yes, it will change frequently
□ Yes, it will change sometimes
□ Not sure
□ No, it will not change
If yes, what is the main reason? (Optional):

(4) If the same sound is given the same prompt without distinguishing scenarios, would you feel that it adds hassle to your life or poses a danger?
(1):When you hear "a horn honking":
A. Indoors (should remind: Can be ignored)
B. On the road (should remind: Watch out for vehicles)
If scenarios are not distinguished, and both scenarios prompt you "Can be ignored", what do you think? (Single choice)
□ Dangerous □ Adds hassle □ Minor impact □ No impact
(2):When you hear "screaming":
A. Amusement park (should remind: Someone nearby is playing)
B. Outdoors (should remind: Look for the source of the scream)
If scenarios are not distinguished, and both scenarios prompt you "Someone nearby is playing", what do you think? (Single choice)
□ Dangerous □ Adds hassle □ Minor impact □ No impact
(3): When you hear "a broadcast":
A. Music broadcast in the square (should remind: Can be ignored)
B. High-speed rail station/airport (should remind: Pay attention to the departure time of the train/flight)
If scenarios are not distinguished, and both scenarios prompt you "Can be ignored", what do you think? (Single choice)
□ Dangerous □ Adds hassle □ Minor impact □ No impact
(4): When you hear "running water":
A. At home (should remind: Please turn off the tap)
B. In the wild (should remind: Watch out for nearby hazardous water sources)
If scenarios are not distinguished, and both scenarios prompt you "Please turn off the tap", what do you think? (Single choice)
□ Dangerous □ Adds hassle □ Minor impact □ No impact
(5): When you hear "thunder":
A. Indoors (should remind: Can be ignored)
B. Outdoors (should remind: Watch out for lightning)

If scenarios are not distinguished, and both scenarios prompt you "Can be ignored", what do you think? (Single choice)
□ Dangerous □ Adds hassle □ Minor impact □ No impact

(5) Scenario description:
Assume the phone recognizes a sound and will give you a prompt. Of the following two prompts, which do you think is more useful?
Prompt 1 (only tells what sound it is): "An alarm is sounding"
Prompt 2 (tells what sound it is + tells you what to do): "An alarm is sounding; suggestion: first check the kitchen/living room; if you smell gas, open the window, shut off the valve, leave, and seek help."
(1):Which prompt would you prefer? (Single choice)
□ Prompt 1: only tells the sound
□ Prompt 2: sound + suggestion
□ Either is fine
□ I don't want either
(2):What is the main reason you chose it? (Multiple choice, select up to 2)
□ I want to know what to do next
□ Only seeing the "sound name", I'm still not sure whether it is serious / whether I should handle it immediately
□ Having suggestions makes me feel more reassured / safer
□ Only stating the sound is enough; I can judge by myself
□ Too many prompts would feel annoying / disruptive
□ Other:
(3):If the phone only prompts the sound (for example, "An alarm is sounding"), would you immediately know what to do? (Single choice)
□ Yes, I would immediately know what to do
□ Sometimes; it depends on the situation
□ Often not; I need to think about it / ask others
□ No; I don't know what this means

A.2 Algorithm

Algorithm 1 Human Speech / Ambient Sound / Safety-Critical Detection Algorithm

Require: $Detector_{result}$: Detect single-frame decision (Silence / Non-silence)
E_k: Energy of each band ($k \in \{0, 1, ..., 6\}$)
$T_E(k)$: Energy threshold of band k
R_k: Energy ratio of band k (%)
E_{min}: Silence threshold
Δ_t: Temporal window for consistency check
Ensure: $Final_{result}$: (0=Silence, 1=Speech, 2=Ambient, 3=Safety-critical)

$E_{total} \leftarrow \sum_k E_k$
if $E_{total} \leq E_{min}$ **then**
  $Final_{result} \leftarrow 0$
else
  if $Detector_{result}$ = Non-silence **then**
    ▷ **Step 1: Safety-critical pattern detection**
    $Siren_{pattern} \leftarrow (R_{400\sim2000} \geq 35) \wedge (TemporalOscillation(\Delta_t) = True)$
    $Alarm_{pattern} \leftarrow (R_{2\sim4kHz} \geq 30) \wedge (PeakRepeat(\Delta_t) = True)$
    if $Siren_{pattern} \vee Alarm_{pattern}$ **then**
      $Final_{result} \leftarrow 3$ ▷ Safety-critical event → deterministic override
    else
      ▷ **Step 2: Human speech verification**
      $Core_1 \leftarrow (E_{400\sim800} \geq T_E) \wedge (R_{400\sim800} \geq 20)$
      $Core_2 \leftarrow (E_{1600\sim3200} \geq T_E) \wedge (R_{1600\sim3200} \geq 15)$
      $Core_{ok} \leftarrow Core_1 \wedge Core_2$
      ▷ **Step 3: Noise filtering**
      $Noise_1 \leftarrow (R_{0\sim200} \leq 10)$
      $Noise_2 \leftarrow (R_{4\sim16kHz} \leq 8)$
      $Noise_{ok} \leftarrow Noise_1 \wedge Noise_2$
      ▷ **Step 4: Auxiliary conditions**
      $Aux_1 \leftarrow (R_{200\sim400} \geq 5)$
      $Aux_2 \leftarrow (R_{800\sim1600} \geq 8)$
      $Aux_{ok} \leftarrow Aux_1 \vee Aux_2$
      if $Core_{ok} \wedge Noise_{ok} \wedge Aux_{ok}$ **then**
        $Final_{result} \leftarrow 1$
      else
        $Final_{result} \leftarrow 2$
  else
    $Final_{result} \leftarrow 2$
  return $Final_{result}$