DO AUDIO LLMS LISTEN BEFORE THEY ACT? DIAGNOSING ACOUSTIC-CONTEXT GATING IN VOICE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio language models can recognize spoken commands and invoke tools, but an agent must first decide whether the acoustic and conversational context warrants action. We introduce VGBENCH, a 1,018-item diagnostic benchmark for action-level addressedness across side-talk, self-talk, and speaker-switch scenarios. Each item uses a shared action space comprising silence, a tool call, and a natural-language answer. Speaker-switch pairs hold the specified words fixed while source, distance rendering, and a temporal boundary define a controlled wearer-to-bystander shift. Six raw Audio LLMs and three training-free adaptations often identify the target tool yet rarely withhold action under this shift; the highest raw switch mute rate is 14%. We then use VOXGATE as a post-training case study. Supervised training mutes 91.3% of switched commands while choosing the correct tool for all nearby wearer commands and text-only controls. An exploratory GRPO stage has similar switch performance; side-talk accuracy rises from 68.4% to 70.9%, and self-talk muting from 52.0% to 60.0%. Factorized controls identify an independent source-change effect, while sensitivity to the far-field manipulation varies across acoustic renderings. The benchmark therefore measures multi-cue acoustic-context gating rather than isolated speaker identity.

1 INTRODUCTION

Voice agents increasingly perform consequential actions: they schedule appointments, send messages, place orders, and control physical systems. A spoken instruction carries both textual content, which specifies an action, and acoustic and conversational context, which indicates whether the utterance is addressed to the assistant. A bystander can utter the same command as the user, and a user can mention a command while thinking aloud. An agent that relies on text alone may recognize the requested operation while making the wrong decision to act.

Existing systems often place keyword spotting, device-directed speech detection, or speaker verification before the agent (Mallidi et al., 2018; Nam et al., 2026). End-to-end Audio LLMs instead expose one policy that can remain silent, answer, or invoke a tool. Existing evaluations usually pair well-formed requests with a response or tool label, so high tool-selection accuracy does not show that acoustic context controls execution. We ask a narrower action-level question: when the specified words are held fixed but the trigger comes from a different acoustic source and scene, does an Audio LLM act or remain silent?

We study this question with VGBENCH, a 1,018-item diagnostic benchmark for *agentic addressedness*. It contains 395 side-talk recordings, 223 self-talk recordings, and 400 paired speaker-switch cases. Every item maps to a shared action space comprising `[Mute]`, a tool call, and a natural-language answer. Side-talk tests conversational attribution within one recording; self-talk tests command-like monologues; speaker-switch reverses the target from a tool call to `[Mute]` while preserving the specified context and trigger words. The switch condition jointly changes trigger source, far-field rendering, and a 600 ms boundary. It therefore tests a controlled wearer-to-bystander source-and-scene shift, not isolated speaker identity.

We evaluate six raw Audio LLMs and three training-free adaptations under the same standardized action contract. The strongest raw switch mute rate is 14%, and the strongest training-free result

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

is 6%, even when target-tool selection is high. Step-Audio-R1.1, for example, selects the target tool on 96% and 97% of same-speaker and text-only controls, respectively, but mutes only 1% of switched commands. These results expose a gap between recognizing command content and using acoustic-pragmatic evidence to control action.

We then use VOXGATE as a post-training case study. Supervised fine-tuning combines the training partition of VGBENCH with answer, abstention, tool-use, and translation examples from WearVox. It accounts for most of the observed gating improvement, muting 91.3% of switched commands while choosing the correct tool for all nearby wearer commands and text-only controls. An exploratory counterfactual-pair GRPO stage yields similar switch muting. In one run, side-talk accuracy rises from 68.4% to 70.9%, self-talk muting from 52.0% to 60.0%, and WearVox overall from 72.14% to 76.30%. These differences do not isolate the effect of pair-aware grouping. Factorized controls show that source change affects action independently, while far-field rendering mutes 60% of same-speaker triggers, as the proximity rule requires. We further use equal-budget cue-balanced SFT to test how these cues affect action.

Our contributions are:

- **Action-level benchmark.** VGBENCH evaluates side-talk, self-talk, and text-matched source-and-scene shifts under one silence, tool, and answer interface.
- **Behavioral diagnosis.** Full-corpus evaluation shows that current systems can recover target tools while failing to condition execution on acoustic and conversational evidence.
- **Post-training case study.** Joint supervised training shows that the measured gate is learnable without collapsing the positive controls; exploratory GRPO and factorized controls characterize the remaining gains and cue dependence.

2 RELATED WORK

Addressedness and voice-agent evaluation. Voice-assistant pipelines traditionally treat "was I spoken to?" as a separate detection problem, using keyword spotting, device-directed-utterance detection, speaker verification, or additional signals such as gaze (Mallidi et al., 2018; Siegert et al., 2022; Zhang & Rekimoto, 2025; Nam et al., 2026). Recent benchmarks bring related decisions into richer agent settings. WearVox uses 3,842 real egocentric multichannel recordings and includes both side-talk rejection and tool calling (Lin et al., 2026). Audio2Tool evaluates spoken tool use across direct, compositional, and acoustically mixed queries (Pahwa et al., 2026). ProVoice-Bench studies when proactive voice agents should intervene or remain dormant (Xu et al., 2026). VGBENCH complements these resources with text-matched action contrasts: the same specified command supports execution in one condition and silence in another. This design connects acoustic-context use directly to action selection rather than treating addressedness as a separate front-end score.

Acoustic and paralinguistic evidence in Audio LLMs. MMAU and MMAU-Pro cover broad audio understanding and reasoning (Sakshi et al., 2025; Kumar et al., 2026). Other benchmarks study prosody and phonology in MMSU (Wang et al., 2026), stress and intonation in WildSpeech-Bench (Zhang et al., 2025b), speaker attributes in SD-Eval (Ao et al., 2024), multi-speaker grounding in MSU-Bench (Sun et al., 2026), and audio-dependent questions in AUDITA (Kabir et al., 2026). Closest to our diagnostic design, VoxParadox constructs linguistic-acoustic conflicts to test whether a model uses paralinguistic cues (Pang et al., 2026). Its targets are perceptual answers; VGBENCH instead asks whether acoustic and pragmatic evidence changes the choice to remain silent, answer, or invoke a tool.

Post-training and inference-time adaptation. GRPO was introduced in DeepSeekMath (Shao et al., 2024) and is now used for audio reasoning post-training (Li et al., 2025; Wen et al., 2025; Zhang et al., 2025a). Omni-R1 shows that text-only fine-tuning can improve audio benchmarks (Rouditchenko et al., 2025), while recent methods explicitly measure audio contribution or penalize late-stage loss of audio attention (He et al., 2026; Xiao et al., 2026). Training-free systems instead structure inference through tool orchestration, acoustic DSP, or chunked listening (Wijngaard et al., 2025; Maben et al., 2025; Xiong et al., 2025; Chiang et al., 2026). We evaluate representative inference-time adaptations as behavioral probes and use post-training to test whether the VGBENCH action boundary is learnable.

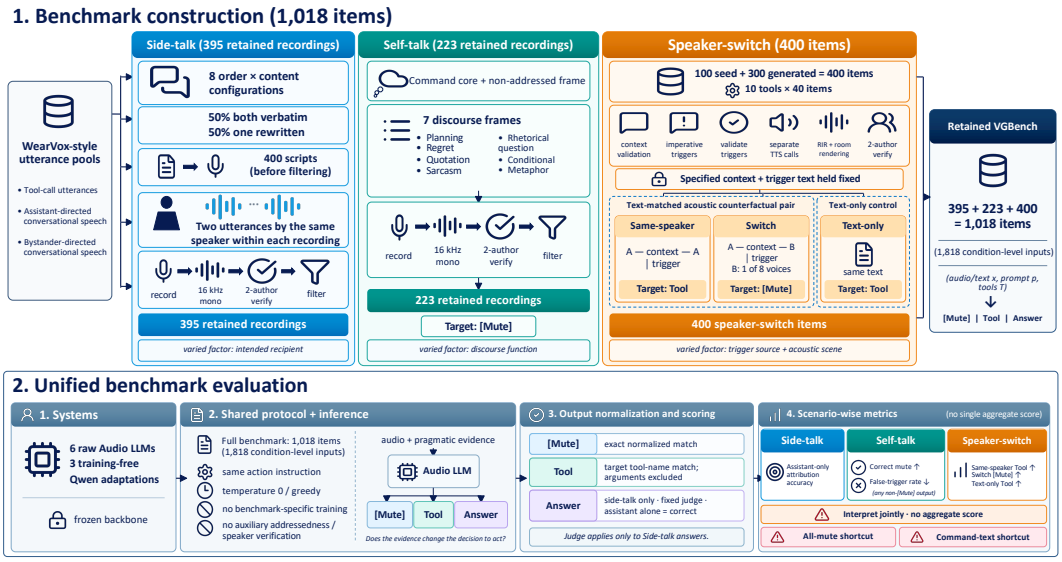

Figure 1: Overview of VGBENCH construction and evaluation. The benchmark contains 1,018 items across side-talk, self-talk, and speaker-switch scenarios. Each speaker-switch item has same-speaker, switch, and text-only conditions. Metrics are interpreted jointly rather than collapsed into one aggregate score.

3 VGBENCH: ACTION-LEVEL ADDRESSEDNESS DIAGNOSTICS

3.1 TASK FORMULATION

An agentic voice assistant receives audio x, a system prompt p, and available tools $\mathcal{T}$. It selects one of three actions: MUTE, TOOL, or ANSWER. MUTE returns the literal token `[Mute]` and executes nothing; TOOL emits one structured call; and ANSWER returns natural language without invoking a tool. We call the joint decision based on textual semantics, intended recipient, and acoustic source *agentic addressedness*. Unlike perceptual classification, this task tests whether contextual evidence changes an executable decision.

3.2 DIAGNOSTIC DESIGN

VGBENCH uses three scenario families to separate complementary failures that are usually collapsed into a single voice-assistant accuracy score. Side-talk asks which utterance within a local exchange licenses action. Self-talk asks whether command-like words are sufficient to trigger an action when the discourse frame marks them as non-instructions. Speaker-switch pairs hold the specified words fixed while reversing the target action under a controlled change in source and scene. The positive controls attached to each family are essential for interpretation: side-talk checks whether the model acts only on the addressed utterance, self-talk tests whether it rejects lexical command shortcuts, and speaker-switch requires action reversal while retaining tool recognition.

Side-talk. Each of 395 recordings contains two consecutive utterances from the same stored speaker, one directed to the assistant and one to a nearby person, with balanced order. Using one voice removes speaker identity as an explanation and tests pragmatic attribution. The agent should act only on the assistant-directed utterance. The scripts also balance tool-like and conversational content, so utterance position or the presence of an obvious command is not by itself a reliable decision rule.

Self-talk. The 223 recordings contain command-like monologues framed as planning, regret, quotation, sarcasm, rhetorical questions, and related discourse forms. Every item targets `[Mute]` even when the words contain a command that maps to an available tool.

Speaker-switch. The 400 counterfactual pairs cover ten consequential tools. For these tool triggers, we adopt a conservative wearable authorization rule: $\text{Tool} \iff$ same authorized source $\wedge$ near-field. A different source or a far-field trigger targets `[Mute]`. In the *same-speaker* condition, source A (the session initiator) speaks the context and near-field trigger, and the target is a tool call. In the *switch* condition, A speaks the context and a bystander source B speaks the same trigger under fixed far-field rendering after a 600 ms boundary, so the target is `[Mute]`. The *text-only* condition uses the same words and retains the tool target. The contrast holds specified text fixed while jointly changing source, distance rendering, and temporal boundary. A far-field trigger from A also targets `[Mute]` under this rule.

3.3 Construction and Splits

Side-talk and self-talk scripts are derived from tool-like, assistant-directed, and bystander-directed seed pools. Their construction balances utterance order and content type and places command cores in varied discourse frames. Commissioned speakers recorded the scripts as natural speech. Recordings are converted to 16 kHz mono PCM WAV and retained after two rounds of annotation, target-label verification, and script-audio consistency checks, yielding 395 side-talk and 223 self-talk recordings.

Speaker-switch uses 100 seed cases and 300 generated cases, balanced across ten consequential tools with 40 cases per tool. Context and trigger segments are synthesized separately using eight distinct voices to simulate speaker changes. The same-speaker condition uses source A for both segments. The switch condition uses source A for context, inserts 600 ms of silence, and renders the trigger from source B with a fixed far-field room transform. Segments are normalized to -20 dBFS with 5 ms fades before 16 kHz mono export. The manifest records voice, duration, normalization, peak, boundary, and room configuration. These controls make the aggregate switch condition reproducible, while also defining its scope: it is a joint source, distance, and boundary intervention.

Raw models and training-free adaptations are evaluated on the full corpus. Post-training uses one approximately 80/20 item split within each scenario, including 320/80 speaker-switch pairs. Only the training partition enters SFT or GRPO, and all reported post-training addressedness scores use the disjoint held-out partition. This protocol establishes item-level separation; it does not establish held-out voice, template-family, or tool-family generalization.

3.4 Evaluation Contract

Every system receives the same tool inventory and instruction to emit exactly one of three action forms. The canonical mute output is `[Mute]`. A tool action is a single structured object, `<|TOOL|>{{"name": name, "params": {...}}}</|TOOL|>`, and an answer contains neither a canonical tool call nor `[Mute]`. We use greedy decoding throughout. Model-specific wrappers preserve required chat templates and response terminators, but evaluation always applies the same normalized parser and target action.

We report scenario-specific metrics jointly. Speaker-switch evaluation combines switch-condition mute rate with target-tool selection in the same-speaker and text-only controls, preventing an always-mute policy from scoring well. Self-talk uses mute rate, and side-talk measures whether the model acts only on the assistant-directed utterance. A correct tool output must be parseable and match the target tool name. Arguments remain in prediction records but are outside the VGBench score, so the metric measures tool-name selection rather than complete executable-call accuracy.

Side-talk tool items use the canonical parser. For free responses, a fixed Qwen3.5-35B-A3B judge assigns one of four labels: assistant-directed, bystander-directed, both, or neither. Only assistant-directed alone is correct. The same system instruction, decoding configuration, parser, judge, prediction schema, and summary procedure are frozen for each comparison block. The current protocol has no reported human-agreement estimate for the judge; this affects the free-response subset rather

Table 1: Action-selection results on the complete VGBENCH corpus. Values are percentages. Speaker-switch reports switch mute rate and target-tool selection in the near-field same-speaker and text-only positive controls. Side-talk combines canonical tool selection and free responses scored by the fixed Qwen3.5-35B judge.

Method	Side-talk correct	Self-talk mute	Switch mute	Same-speaker select.	Text-only select.
Raw Audio LLMs					
Qwen3-Omni	39.0	57.0	1	77	81
Nemotron-Omni	32.7	28.7	0	84	85
Gemini-3.8-flash	32.4	68.6	7	80	71
Kimi-Audio-7B	38.2	30.5	**14**	53	44
Step-Audio-R1.1	41.0	22.4	1	**96**	97
Audio Flamingo 3	36.0	1.0	0	0	0
Training-free adaptations					
Qwen + AURA adaptation	**46.6**	63.2	0	95	**99**
Qwen + TwS adaptation	23.0	**85.7**	6	90	82
Qwen + SHANKS adaptation	17.5	79.8	1	1	81

than the rule-scored mute and tool decisions. Appendix A records the result artifacts and scorer provenance.

4 OFF-THE-SHELF ACTION POLICIES

4.1 SETUP

We evaluate six raw Audio LLMs: Qwen3-Omni-30BXu et al. (2025), Nemotron-3-Nano-Omni-30BDeshmukh et al. (2026), Gemini-3.8-flashDoshi & Popa (2026), Kimi-Audio-7B (Ding et al., 2025), Step-Audio-R1.1 (Tian et al., 2025), and Audio Flamingo 3 (Ghosh et al., 2026). We also test local, method-style adaptations of AURA, Thinking with Sound (TwS), and SHANKS on the Qwen backbone (Maben et al., 2025; Xiong et al., 2025; Chiang et al., 2026). These implementations probe inference-time prompting, acoustic tools, and chunked reasoning; they are not claimed as official reproductions. All systems use temperature zero, the same action instruction, and the same scorer. Model-specific normalization maps native outputs to the canonical parser, so the raw comparison reflects both action behavior and compatibility with the standardized interface.

Tool recognition does not imply acoustic-context gating. Step-Audio selects the target tool on 96% and 97% of same-speaker and text-only controls, respectively, yet mutes only 1% of switched commands. Kimi has the highest raw switch mute rate at 14%, but selects the target tool on only 53% and 44% of the two controls. These paired measurements expose failure to change the action under a source-and-scene shift, rather than a general inability to recognize tools.

Systems fail through different shortcuts. Qwen preferentially acts on the first side-talk utterance, Gemini favors the later utterance, and Step often responds to both. Audio Flamingo 3 receives zero canonical tool-selection credit because it emits a different bracketed action language. This format mismatch limits cross-model comparison of positive-control scores, but it does not explain its 1% self-talk mute rate or zero switch mute rate. The standardized interface therefore diagnoses deployable action behavior, not a format-invariant latent capability.

Inference-time adaptations change the error profile. AURA improves side-talk from 39.0% to 46.6% and self-talk from 57.0% to 63.2%, but mutes no switched commands. TwS raises self-talk muting to 85.7% while side-talk falls to 23.0%, largely through overuse of silence. SHANKS also raises self-talk muting, but its chunked output often violates the action grammar, leaving 1% same-speaker tool selection. Acoustic evidence may appear in intermediate reasoning without controlling the final action, an action-level counterpart to the utilization gap studied by VoxParadox.

5 CAN ACOUSTIC-CONTEXT GATING BE LEARNED?

We use VOXGATE as a post-training intervention rather than as evidence for a new speaker-identification mechanism. Supervised fine-tuning is the primary intervention; an exploratory GRPO stage tests whether paired rollouts preserve or improve the learned action boundary. Both stages use the same action contract as the benchmark.

5.1 SUPERVISED JOINT TRAINING

The model outputs $y \in \{\text{MUTE}, \text{TOOL}, \text{ANSWER}\}$. We train a LoRA adapter on a mixture of the VGBENCH training partition and WearVox answer, abstention, tool-use, and translation examples. The addressedness data include side-talk attribution, self-talk mute targets, and speaker-switch counterfactuals. In each switch pair, matched words support a tool call for the same-source near-field condition and `[Mute]` for the bystander condition; text-only examples retain the tool target.

We use Qwen3-Omni-30B-A3B-Instruct with LoRA rank 32 and alpha 64 on attention q, k, v, and o projections across 48 layers. The audio encoder and aligner are frozen. Training uses three epochs, learning rate 10^{-4}, effective batch size 32, and eight H200 GPUs.

For the cue-balanced SFT control, we replace only 1,212 speaker-switch audio training slots in the 5,748-row mixture. Other task proportions, 540 optimization steps, and the LoRA configuration remain fixed. Within both tool and `[Mute]` labels, near/far rendering is crossed with 0/600 ms separation and balanced across the four combinations. The mixture includes far-field authorized-source/tool rows, making this a cue-control experiment rather than a proximity-gate training protocol. Appendix H details the construction.

5.2 EXPLORATORY COUNTERFACTUAL-PAIR GRPO

For a speaker-switch pair $q = (x^+, x^-)$, Stage II samples $K/2$ completions from each condition and places them in one rollout group. Here x^+ is the same-speaker condition with a tool target, and x^- is the switched condition with a mute target. Grouping matched text with opposite actions makes reward normalization depend on whether the policy separates the counterfactual pair, rather than on unrelated prompt difficulty.

Let $r(y_{q,i}, x_{q,i})$ be the rule-based reward for completion i. We compute

$$A_{q,i} = \frac{r(y_{q,i}, x_{q,i}) - \mu_q}{\sigma_q + \epsilon}, \qquad \mu_q = \frac{1}{K}\sum_{i=1}^{K} r(y_{q,i}, x_{q,i}), \tag{1}$$

where μ_q and σ_q are calculated across both sides of the pair. Only groups with nonzero reward variance contribute an update. This removes groups in which every sampled action receives the same reward and hence provides no within-group policy-gradient signal.

A correct mute or matching tool receives $+1$, and an incorrect action receives -1; malformed tool syntax incurs an additional -0.2. For answer targets, nonempty natural language receives $+0.5$, mute receives -1, and empty or tool-only output receives -0.5. WearVox examples retain task-specific rewards for answer, abstention, structured tool use, and translation. We track rewards by scenario because an aggregate curve can hide a saturated positive control or a failed mute boundary. The post-training mixture, frozen components, and decoding contract are shared with SFT unless stated otherwise.

This stage tests whether paired optimization preserves or improves the learned action boundary. The experiment has no unpaired-GRPO or repeated-seed control, so it does not isolate pair-aware grouping as the cause of an SFT-to-GRPO difference. Appendix D records the full implementation configuration.

6 POST-TRAINING RESULTS

The base, supervised, and GRPO models use the same system instruction, greedy decoding, canonical action parser, and fixed Qwen3.5-35B-A3BQwen Team (2026) side-talk judge. All addressedness results use the disjoint held-out partition. Task-only SFT and GRPO controls use WearVox

Table 2: Addressedness results on the fixed 20% VGBENCH test split. Values are percentages. Switch mute is reported with near-field same-speaker and text-only target-tool selection to expose all-mute behavior.

Model or training setting	Self-talk mute	Side-talk correct	Same-speaker select.	Switch mute	Text-only select.
Base Qwen	50.0	53.2	81.3	0.0	87.5
VOXGATE supervised training	52.0	68.4	**100.0**	91.3	**100.0**
VOXGATE + GRPO	**60.0**	**70.9**	**100.0**	**92.5**	**100.0**

without VGBENCH examples. Downstream evaluation follows the fixed 384-example WearVox protocol.

Supervised training accounts for most of the switch result. It mutes 73 of 80 held-out switched commands while selecting the target tool for every near-field same-speaker and text-only control. GRPO changes the switch result to 74 of 80, raises self-talk muting from 52.0% to 60.0%, and raises side-talk accuracy from 68.4% to 70.9%. These are descriptive differences from one training run. They show that GRPO preserves the SFT gate, but do not establish that pair-aware grouping caused the improvement.

The aggregate switch score combines several cues. Table 3 factorizes trigger source, distance, and the 600 ms boundary on the same 80 commands. With near-field rendering and a fixed gap, changing only the trigger source raises mute rate from 1.25% to 50.0% for SFT and to 53.75% for GRPO. Changing a same-speaker trigger from near-field to far-field raises muting to 60.0% for both models, appropriate under the proximity rule. The gap alone changes same-speaker near-field muting by only 1.25 points. The 91.3–92.5% main switch result therefore combines source and distance dependence.

Table 3: Mute rates (%) on the original 80-command cue validity WAVs. Same/near targets a tool; same/far and different-speaker conditions target `[Mute]`. The last row is the original switch condition.

Source	Field	Gap	Base	SFT	GRPO
Same	Near	0	7.50	0.00	0.00
Same	Near	600	5.00	1.25	1.25
Same	Far	600	6.25	60.00	60.00
Different	Near	0	6.25	35.00	31.25
Different	Near	600	7.50	50.00	53.75
Different	Far	600	0.00	91.25	92.50

6.1 CUE-BALANCED SFT AS A CUE CONTROL

We call the supervised model in Tables 2 and 3 original SFT; cue-balanced SFT is a separate SFT-only run, with no subsequent GRPO stage. The original SFT learned from switch examples in which bystander speech, far-field rendering, and a 600 ms gap occurred together. We retrain with the same budget after balancing distance and gap within both action labels, then test all eight source, distance, and gap combinations on 80 held-out commands.

Table 4: Target-action accuracy (%) under the proximity rule in the scene-consistent eight-condition test, original SFT → cue-balanced SFT. Each cell has $n = 80$; equal-weight accuracy is 39.22% → 66.88%. Same-speaker triggers target a tool when near and `[Mute]` when far; different-speaker triggers target `[Mute]`. Complete mute/tool rates appear in Appendix H.

Field, gap	Same: Tool near, `[Mute]` far	Different: `[Mute]`
Near, 0 ms	97.50 → 97.50	18.75 → 68.75
Near, 600 ms	100.00 → 98.75	18.75 → 85.00
Far, 0 ms	2.50 → 3.75	23.75 → 80.00
Far, 600 ms	6.25 → 5.00	46.25 → 96.25

Direct retesting on the original validity WAVs raises same/far/600 muting from 60.00% to 90.00% and different/near muting from 35.00%/50.00% to 66.25%/82.50% at 0/600 ms (Appendix G, Table 12). The eight-condition study instead uses scene-consistent far-field audio, rendering the wearer's context and trigger together; it yields only 3.75%/5.00% muting of same/far triggers. These are different WAVs, so distance rejection is scene dependent.

6.2 Downstream Task Retention

The addressedness objective should not improve muting by erasing the model's ability to answer, use tools, or translate. We therefore evaluate the post-trained models on the fixed 384-example WearVox protocol. It contains 59 closed-book answer examples, 55 grounded-answer examples, 58 abstention examples, 112 tool-use examples, and 100 live-translation examples. The A score aggregates the two answer subsets.

Table 5: Held-out WearVox task retention for the original SFT/GRPO training route (cue-balanced SFT is reported separately above). A aggregates closed-book and grounded answers; B, C, and D denote abstention, tool use, and live translation. Counts are shown in the column headers. D is the mean judged translation score ($\times 100$); Overall thresholds each translation item at 0.85 before aggregating all 384 binary outcomes.

Model	A ($n = 114$)	B ($n = 58$)	C ($n = 112$)	D ($n = 100$)	Overall ($n = 384$)
Base Qwen	28.11	96.60	55.40	65.30	45.57
WearVox-only post-training					
SFT	33.33	94.83	85.71	84.80	63.28
SFT + GRPO	42.98	98.28	86.61	83.43	67.71
Joint VoxGate post-training					
VoxGate supervised training	37.71	91.40	90.20	89.50	72.14
VoxGate + GRPO	42.10	**100.00**	**92.00**	**89.70**	**76.30**

Joint training retains downstream task performance. Task-only SFT and GRPO reach 63.28% and 67.71% overall, while joint supervised training and joint GRPO reach 72.14% and 76.30%. The joint models improve tool use and translation relative to task-only controls; joint GRPO also attains 100.00% abstention accuracy. The answer aggregate remains lower than the other task families for every system, including the base model, and the joint GRPO answer score is close to the task-only GRPO score. These results show that the learned benchmark gate does not come from a general collapse of retained tasks. WearVox retention does not, however, measure transfer to natural addressedness interactions because its downstream categories and scorer serve a different evaluation purpose.

7 Discussion

Action selection is distinct from command recognition. The strongest positive-control results coexist with near-zero switch muting for off-the-shelf models. Under the standardized interface, an Audio LLM can recover a command and select its tool without using acoustic context to decide whether the action should execute. VGBench makes this gap observable by pairing positive and negative actions with the same specified words.

The learned gate uses source and proximity cues. At fixed near-field rendering and a 600 ms gap, changing the trigger source raises muting by 48.75 points for SFT and 52.50 for GRPO. At fixed source and gap, far-field rendering raises it by 58.75 points for both. Under the proximity rule, the original SFT's 60% same/far muting is appropriate, while its 35–50% different/near muting leaves many bystander triggers executable. The gap amplifies source changes but is not a standalone mute rule.

Paired reporting separates recognition from policy errors. A mute score alone rewards a degenerate policy that never acts, while tool accuracy alone rewards a policy that executes every recognized command. The near-field same-speaker and text-only controls expose false rejection; different-speaker conditions expose false execution. The factorized study changes one cue at a time while preserving the command and matching-tool identity, revealing near-field bystander errors that an aggregate score would conceal.

Implications for voice-agent design. An action gate needs both proximity and source evidence: distance alone cannot reject a nearby bystander. Evaluation should report false execution and false rejection separately.

The scenario families test different information pathways. Side-talk tests recipient attribution, self-talk tests whether discourse cancels command force, and speaker-switch changes source and scene while holding words fixed. Performance on one family does not imply performance on the others, so we report them separately.

What post-training establishes. Supervised training learns the benchmark's conditional action mapping while preserving near-field and text-only tool controls and WearVox tasks. The cue study shows why an aggregate action reversal cannot identify the rule: source and distance can each change the decision.

Cue balancing improves near-bystander suppression but is scene dependent. The equal-budget SFT control crosses distance and gap within both training labels, including far-field wearer/tool rows; it probes cue sensitivity rather than training the proximity policy. On the original validity WAVs, it raises different/near muting from 35.00%/50.00% to 66.25%/82.50% and same/far muting from 60.00% to 90.00%. On scene-consistent eight-condition audio, proximity-rule accuracy rises from 39.22% to 66.88%, but same/far muting remains only 3.75%/5.00%. Original speaker-switch muting falls from 73/80 to 70/80, and WearVox tool-call accuracy falls from 90.18% to 83.04%. The matched-WAV gains do not establish a reliable proximity gate across acoustic scenes.

Current evidence boundaries. Results use one split and one run per setting. The balanced test varies wearer voice but not room, template, natural bystander speech, or open-set speakers. Side-talk uses an LLM judge without reported human agreement; VGBENCH tool accuracy checks names rather than arguments.

8 CONCLUSION

VGBENCH shows that Audio LLMs often recognize spoken commands without reliably choosing the intended action across 1,018 controlled items. Supervised training learns much of this benchmark action boundary, while an exploratory GRPO stage also adds small gains in one run. Factorized controls identify an independent source-change effect, but sensitivity to far-field rendering varies across acoustic scenes. Cue-balanced SFT improves near-bystander suppression without establishing a reliable cross-scene proximity gate. These results highlight a gap between understanding what was said and deciding whether acoustic and conversational context warrants action, and motivate evaluating voice agents at the level of executable action rather than command recognition alone.

AI USE STATEMENT

Generative AI tools assisted language editing of portions of the manuscript, dataset construction, and user simulation, and served as judging components. The authors reviewed, executed, and verified AI-assisted code and reviewed every edited passage. The authors take responsibility for the final text, artifacts, and claims.

ETHICS STATEMENT

Side-talk and self-talk include task-focused speech recorded for this study; speaker-switch audio is synthesized. The supplementary package contains scripts, text-side annotations, and sample clips, without full recordings. Incorrectly acting on withdrawn or bystander requests can cause harm in high-stakes assistant settings. Tool calls in this benchmark are scored offline.

REPRODUCIBILITY STATEMENT

The supplementary package provides text-side records for all 1,018 benchmark items, construction manifests, generation and evaluation scripts, a judge client, and SFT/GRPO code and recipes. It includes representative examples for each task family, with three audio clips per scenario. The complete audio corpus exceeds the OpenReview supplementary-material size limit and cannot be distributed through the available anonymous review-time hosting options without compromising the double-blind submission setup. We therefore provide representative audio during review and will release the full dataset, including all audio files, upon acceptance. Evaluations require the audio at the paths specified in the manifests and compatible model and judge servers. Speaker-switch synthesis additionally requires an external TTS model and reference voices.

REFERENCES

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *Advances in Neural Information Processing Systems*, 37:56898–56918, 2024.

Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. Shanks: Simultaneous hearing and thinking for spoken language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8951–8972, 2026.

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, et al. Nemotron 3 nano omni: efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Tulsee Doshi and Raluca Ada Popa. Introducing gemini 3.8 flash and 3.8 flash cyber. `https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/`, September 2026. Google Blog.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886, 2026.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, et al. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. In *International Conference on Learning Representations*, volume 2026, pp. 110411–110432, 2026.

Tasnim Kabir, Dmytro Kurdydyk, Aadi Palnitkar, Liam Dorn, Ahmed Haj Ahmed, and Jordan Lee Boyd-Graber. Audita: A new dataset to audit humans vs. ai skill at audio qa. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25922–25951, 2026.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22688–22697, 2026.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025.

Zhaojiang Lin, Yong Xu, Kai Sun, Jing Zheng, Yin Huang, Surya Appini, Krish Narang, Renjie Tao, Ishan Jain, Siddhant Arora, et al. Wearvox: An egocentric multichannel voice assistant benchmark for wearables. In *International Conference on Learning Representations*, volume 2026, pp. 66555–66574, 2026.

Leander Melroy Maben, Gayathri Ganesh Lakshmy, Srijith Radhakrishnan, Siddhant Arora, and Shinji Watanabe. Aura: agent for understanding, reasoning, and automated tool use in voice-driven tasks. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–4. IEEE, 2025.

Sri Harish Mallidi, Roland Maas, Kyle Goehner, Ariya Rastrow, Spyros Matsoukas, and Björn Hoffmeister. Device-directed utterance detection. *arXiv preprint arXiv:1808.02504*, 2018.

KiHyun Nam, Jungwoo Heo, Siu Bae, Ha-Jin Yu, and Joon Son Chung. Speakerllm: A speaker-specialized audio-llm for speaker understanding and verification reasoning. *arXiv preprint arXiv:2605.15044*, 2026.

Ramit Pahwa, Apoorva Beedu, Parivesh Priye, Rutu Gandhi, Saloni Takawale, Aruna Baijal, and Zengli Yang. Audio2tool: Speak, call, act–a dataset for benchmarking speech tool use. *arXiv preprint arXiv:2604.22821*, 2026.

Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. Do audio llms listen or read? analyzing and mitigating paralinguistic failures with voxparadox. *arXiv preprint arXiv:2605.27772*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-r1: Do you really need audio to fine-tune your audio llm? In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2025.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Ingo Siegert, Norman Weißkirchen, and Andreas Wendemuth. Acoustic-based automatic addressee detection for technical systems: A review. *Frontiers in Computer Science*, 4:831784, 2022.

Zhaokai Sun, Shuai Wang, Zhennan Lin, Chengyou Wang, Dehui Gao, Yuang Cao, Chunjiang He, Pan Zhou, and Lei Xie. Msu-bench: Towards speaker-centric understanding in conversational multi-speaker scenarios. *arXiv preprint arXiv:2606.22868*, 2026.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, et al. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2026, pp. 31374–31410, 2026.

Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. Sari: Structured audio reasoning via curriculum-guided reinforcement learning. *arXiv preprint arXiv:2504.15900*, 2025.

Gijs Wijngaard, Elia Formisano, Michel Dumontier, and Jenia Jitsev. Audiotoolagent: An agentic framework for audio-language models. *arXiv preprint arXiv:2510.02995*, 2025.

Cihan Xiao, Yiwen Shao, Chenxing Li, Xiang He, Zhenwen Liang, Steve Yves, Sanjeev Khudanpur, and Liefeng Bo. Escape the language prior: Mitigating late-stage modality collapse in audio reasoning via modality-aware policy optimization. *arXiv preprint arXiv:2605.27741*, 2026.

Zhen Xiong, Yujun Cai, Zhecheng Li, Junsong Yuan, and Yiwei Wang. Thinking with sound: Audio chain-of-thought enables multimodal reasoning in large audio-language models. *arXiv preprint arXiv:2509.21749*, 2025.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Ke Xu, Yuhao Wang, and Yu Wang. From reactive to proactive: Assessing the proactivity of voice agents via provoice-bench. *arXiv preprint arXiv:2604.15037*, 2026.

Haoyu Zhang, Jiaxian Guo, Dong Yang, Yusuke Iwasawa, and Yutaka Matsuo. Aqa-ttrl: Self-adaptation in audio question answering with test-time reinforcement learning. *arXiv preprint arXiv:2510.05478*, 2025a.

Linhao Zhang, Jian Zhang, Bokai Lei, Chuhan Wu, Aiwei Liu, Wei Jia, and Xiao Zhou. Wildspeech-bench: Benchmarking end-to-end speechllms in the wild. *arXiv preprint arXiv:2506.21875*, 2025b.

Qing Zhang and Jun Rekimoto. Look and talk: Seamless ai assistant interaction with gaze-triggered activation. In *Proceedings of the Augmented Humans International Conference 2025*, pp. 418–421, 2025.

A Evaluation Contract and Reproducibility Record

Each result is tied to a frozen benchmark manifest, system instruction, decoding configuration, action parser, side-talk judge, prediction file, and summary file. Raw Audio LLMs and training-free adaptations are evaluated on all retained VGBench items. Post-trained systems are trained on an approximately 80% partition and reported only on the disjoint held-out partition. We use greedy decoding throughout. The full corpus contains 395 side-talk recordings, 223 self-talk recordings, and 400 speaker-switch pairs. Each speaker-switch pair has single, switch, and text-only conditions; the held-out post-training speaker-switch set therefore contains 80 cases per condition.

Table 6: Result blocks and evaluation records. "Full" denotes the complete retained VGBench corpus, while "held-out" denotes the fixed post-training test partition.

Result block	Systems	Split	Decoding	Scoring record
Raw Audio LLMs	Six native systems	Full	Greedy	Canonical parser; Qwen3.5 side-talk judge
Training-free methods	AURA, TwS, SHANKS	Full	Greedy	Same action contract and judge
VoxGate gating	Base, SFT, GRPO	Held-out	Greedy	Same instruction, parser, and judge
WearVox retention	Base and post-training controls	384 items	Greedy	Official A/B/C/D scorer and translation judge

Table 7: Complete full-corpus raw and training-free results. The corpus contains 395 side-talk recordings, 223 self-talk recordings, and 400 speaker-switch pairs. All values are percentages.

Method	Side-talk	Self-talk mute	Switch mute	Single select.	Text-only select.
Raw Audio LLMs					
Qwen3-Omni	39.0	57.0	1	77	81
Nemotron-Omni	32.7	28.7	0	84	85
Gemini-3.8-flash	32.4	68.6	7	80	71
Kimi-Audio-7B	38.2	30.5	14	53	44
Step-Audio-R1.1	41.0	22.4	1	96	97
Audio Flamingo 3	36.0	1.0	0	0	0
Training-free adaptations					
Qwen + AURA	46.6	63.2	0	95	99
Qwen + TwS	23.0	85.7	6	90	82
Qwen + SHANKS	17.5	79.8	1	1	81

B Action Grammar and Scoring

The canonical mute output is `[Mute]`. A canonical tool action is a single structured object, `<|TOOL|>{{"name": name, "params": {...}}}</|TOOL|>`. A natural-language answer contains neither a canonical tool call nor `[Mute]`. The VGBench parser normalizes the mute token and checks that a tool output contains one parseable object with the target tool name. Tool arguments are retained in prediction logs but are not included in VGBench tool-selection accuracy; they are evaluated by the downstream WearVox tool-use task.

For side-talk free-response examples, the fixed judge selects one of four labels: assistant-directed utterance, bystander-directed utterance, both, or neither. Only the first label is correct. Speaker-switch mute is always interpreted with same-speaker and text-only tool-selection controls, since a system that always mutes cannot be a valid action gate.

C Complete Results for Raw and Training-Free Systems

Table 1 reports the compact comparison. Table 7 adds the scenario denominators and separates the raw and training-free blocks. The values are the same measurements as the main table, rather than a second selection pass. Side-talk combines tool items scored by the canonical parser and response items scored by the fixed side-talk judge.

The three training-free methods use no weight update. AURA adds a ReAct-style attribution step before final action selection. TwS supplies a bounded acoustic analysis protocol with local voice, pitch, energy, and spectral tools. SHANKS uses chunked listening and intermediate notes. These are local implementations of the respective method ideas and are not claimed as official full-pipeline reproductions.

Kimi-Audio receives the system instruction in its first user message because its native template does not retain a system role. Step-Audio uses its required response terminators and long-audio handling. Audio Flamingo 3 emits a native bracketed action language, which does not satisfy the canonical tool contract. This explains its strict tool-selection zeros, but not its addressedness errors: it mutes only 1% of self-talk items and no switched commands.

D VoxGate Implementation Details

Equation 1 defines the paired normalized advantage, and Section 5 specifies the action rewards. The implementation records per-family reward distributions and the number of nonzero-variance rollout groups so that addressedness updates can be distinguished from retained-task updates.

Table 8: VoxGate implementation configuration.

Component	Configuration
Base model	Qwen3-Omni-30B-A3B-Instruct
SFT adaptation	LoRA on all 48-layer attention q, k, v, o projections; rank 32, alpha 64
Frozen modules	Audio encoder and aligner
SFT optimization	Three epochs, learning rate 10^{-4}, effective batch size 32, eight H200 GPUs
SFT mixture	WearVox answer, abstention, tool-use, and translation data plus the approximately 80% VGBENCH training partition
GRPO mixture	The same downstream task mixture and addressedness examples; pair-aware grouping for speaker-switch
Rollouts	$K/2$ completions for the act and mute sides of each speaker-switch pair
Action rewards	Mute, target tool, answer, and task-specific WearVox rewards
Evaluation	Greedy decoding; fixed instruction, parser, and side-talk judge

Table 9: Speaker-switch construction controls.

Variable	Controlled construction
Specified text	Matched context and trigger across single, switch, and text-only conditions
Action target	Tool in single/text-only; `[Mute]` in switch
Voice synthesis	Eight distinct voices simulate speaker changes
Temporal boundary	600 ms silence before the switch trigger
Acoustic scene	Far-field room rendering on the B trigger only
Coverage	400 cases, 10 tools, 40 cases per tool

E DOWNSTREAM WEARVOX EVALUATION

Table 5 reports the complete 384-example WearVox breakdown, including task denominators. Although WearVox-only and joint-model outputs are archived separately, all Table 5 rows were evaluated on the same 384 items with identical task prompts, judge model and prompt, and A/B/C/D scoring code.

F CONSTRUCTION AND QUALITY CONTROLS

Side-talk scripts are balanced over utterance order and tool-like versus conversational content, producing eight order/content-pairing configurations. Self-talk items place a command core in discourse frames including planning, regret, quotation, sarcasm, and rhetorical questions. Recordings are converted to 16 kHz mono PCM WAV and retained only after annotation, label, and script-audio consistency checks.

For speaker-switch, 400 cases are balanced across ten consequential tools (40 cases per tool). Eight distinct voices are used to synthesize speaker changes. The single condition renders source A for both context and trigger. The switch condition renders source A for context, inserts 600 ms of silence, and renders the trigger from source B. The B trigger receives fixed far-field room rendering. Audio is normalized to -20 dBFS per segment with 5 ms fades and 16 kHz mono PCM output. The synthesis manifest records voice, duration, normalization, peak, boundary, and room configuration for each case.

Table 10: Controlled conditions in the speaker-switch validity set. Each condition contains the same 80 held-out commands. Targets follow the proximity rule.

Trigger source	Trigger field	Gap	Target
Same speaker	Near	0 ms	Tool
Same speaker	Near	600 ms	Tool
Same speaker	Far	600 ms	`[Mute]`
Different speaker	Near	0 ms	`[Mute]`
Different speaker	Near	600 ms	`[Mute]`
Different speaker	Far	600 ms	`[Mute]`

Table 11: Speaker-switch cue validity results on the frozen 80-command test set. Cells report mute rate (M) and matching-tool selection rate (T) in percent. Under the proximity rule, same/near targets T and all other audio rows target M. The final row is the original switch condition from Table 2; its non-target tool rate was not retained in the reported summary and is marked with "–".

Source	Field	Gap	Target	Base Qwen		VoxGate SFT		VoxGate GRPO	
				M	T	M	T	M	T
Same	Near	0 ms	Tool	7.50	65.00	0.00	100.00	0.00	100.00
Same	Near	600 ms	Tool	5.00	68.75	1.25	98.75	1.25	98.75
Same	Far	600 ms	`[Mute]`	6.25	73.75	60.00	40.00	60.00	40.00
Different	Near	0 ms	`[Mute]`	6.25	75.00	35.00	65.00	31.25	68.75
Different	Near	600 ms	`[Mute]`	7.50	71.25	50.00	50.00	53.75	46.25
Different	Far	600 ms	`[Mute]`	0.00	–	91.25	–	92.50	–

G Speaker-Switch Cue Validity Study

We construct an evaluation-only validity set from the 80 frozen held-out speaker-switch commands. The set is excluded from training, prompt tuning, checkpoint selection, and model selection. For each command, we synthesize the wearer context, wearer trigger, and bystander trigger once and reuse those source segments across six conditions. This construction holds text, tool target, TTS configuration, sample rate, segment-level loudness, and scoring fixed while varying speaker relation, trigger distance, and temporal separation. Far-field segments are normalized after room rendering so that distance does not introduce a systematic level difference. All 480 generated files pass format, hash, peak, gap, and segment-level loudness checks.

Table 11 reports mute and matching-tool selection rates separately for every condition. The raw base model changes little across the six conditions. Both post-trained models respond to source change under near-field rendering, but they also use distance as a strong gating cue. The 600 ms boundary alone has little effect on same-speaker commands and instead amplifies muting when the trigger source changes.

The paired contrasts clarify the contribution of each cue. At fixed near-field rendering and 600 ms separation, changing the trigger speaker increases mute rate by 48.75 points for SFT and 52.50 points for GRPO. At fixed speaker identity and 600 ms separation, the far-field transform increases mute rate by 58.75 points for both models. Removing the gap from a different near-field source reduces mute rate by 15.0 points for SFT and 22.5 points for GRPO. In contrast, adding the gap to the same near-field wearer changes mute rate by only 1.25 points and preserves 98.75% target-tool selection. The original speaker-switch condition therefore combines source, distance, and turn-boundary evidence, while the same-speaker control shows that the temporal pause is not learned as a standalone mute rule.

Table 12: Cue-balanced SFT on the original six-condition validity WAVs ($n = 80$ per condition), using the same prompts, temperature 0, parser, and scorer as the original validity study. M is `[Mute]` rate and T is matching-tool selection rate (%). These WAVs differ from the scene-consistent eight-condition set in Table 13.

Source	Field	Gap	M	T
Same	Near	0 ms	2.50	97.50
Same	Near	600 ms	2.50	97.50
Same	Far	600 ms	90.00	10.00
Different	Near	0 ms	66.25	33.75
Different	Near	600 ms	82.50	17.50
Different	Far	600 ms	93.75	5.00

H CUE-BALANCED SFT CONTROL

The cue-balanced run changes only the 1,212 frozen-train speaker-switch audio slots in the 5,748-row original SFT mixture: 306 wearer/tool and 906 bystander/`[Mute]` slots. The 286 speaker-switch text-only tool rows and all other task rows remain unchanged. The four near/far by 0/600 ms combinations receive 77, 77, 76, and 76 wearer tool rows, and 227, 227, 226, and 226 bystander mute rows, respectively. The mixture includes far-field authorized-source/tool examples and serves as a cue-control experiment rather than proximity-policy training. Training retains three epochs, 540 steps, and the LoRA settings in Table 8. The 320 training commands and 80 test commands are disjoint; all eight variants of a command stay within its split. Training wearer voices are `M000`, `M008`, `mom`, `dad`, `kid_m`, and `grandma`; test wearer voices are `teen_f`, `phone_caller`, and `tv_host`.

Each command reuses one synthesized wearer context, wearer trigger, and bystander trigger across its eight variants. All speech segments share the same text, TTS checkpoint, 16 kHz mono PCM format, and segment-level loudness normalization. Far-field room rendering and low-pass filtering affect both the wearer context and trigger in a far-field variant; rendered voiced segments are renormalized. The 0/600 ms contrast changes only the inserted silence. In the earlier six-condition cue-validity study, far-field rendering was a trigger-only intervention. Consequently, even with the same 80 command texts, the old and new conditions are different WAVs and their cells should not be pooled or used as matched before/after measurements.

Table 13: Complete eight-condition results on scene-consistent audio. Each row contains 80 held-out commands. M is `[Mute]` rate and T is matching-tool selection rate (%). Same/near targets T; same/far and different-speaker rows target M under the proximity rule.

Relation	Field	Gap	Original SFT		Cue-balanced SFT	
			M	T	M	T
Same	Near	0 ms	2.50	97.50	2.50	97.50
Same	Near	600 ms	0.00	100.00	1.25	98.75
Same	Far	0 ms	2.50	97.50	3.75	96.25
Same	Far	600 ms	6.25	93.75	5.00	95.00
Different	Near	0 ms	18.75	81.25	68.75	31.25
Different	Near	600 ms	18.75	81.25	85.00	15.00
Different	Far	0 ms	23.75	76.25	80.00	20.00
Different	Far	600 ms	46.25	53.75	96.25	3.75