CUEDUET: DISTILLING WHAT TO SAY AND HOW TO SPEAK FOR PARALINGUISTIC-AWARE SPOKEN DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech language models have made substantial progress in spoken interaction, yet adapting response content and vocal delivery to paralinguistic cues remains challenging. Teacher-guided adaptation can transfer cue-informed dialogue behavior, but its supervision mixes general response capabilities with the local adjustments needed for paralinguistic cues. Moreover, appropriate wording alone does not ensure that the generated speech conveys a suitable vocal style. We propose CueDuet, a dual-stream on-policy distillation framework that addresses these complementary needs through teacher supervision on student-generated responses. For the text stream, a single teacher evaluates the same response prefix from cue-aware and cue-agnostic perspectives; divergence between its predictions determines where to emphasize supervision for paralinguistic adaptation. For the speech stream, a speech teacher adapted on expressive text–speech pairs conditions on the student's reply, guiding speech generation toward faithful content and an appropriate style. The evaluation covers emotion- and age-sensitive dialogue across three benchmarks and two architecturally distinct speech language models, comparing offline distillation and standard on-policy distillation baselines. Across both backbones, CueDuet improves cue-responsive content and vocal delivery over offline and standard on-policy distillation, with particularly strong gains in cue-grounded safety. Text-loss analyses show that CueDuet consistently concentrates supervision on cue-responsive positions throughout training. Project page and audio examples: `anonymous demo`.

1 INTRODUCTION

Speech is a natural medium of communication that conveys meaning through both language and vocal expression. Recent advances in speech language models (SLMs) have brought speech understanding and generation into closer integration with language reasoning, supporting cross-modal instruction following, low-latency responses, and real-time full-duplex interaction (Zhang et al., 2023; Fang et al., 2025; Défossez et al., 2024; Wang et al., 2025b). Effective interaction, however, also depends on paralinguistic information beyond the transcript: emotion, prosody, and speaker attributes can shape both what a system should say and how it should speak (Figure 1). Recent research addresses both the perception and interpretation of paralinguistic cues (Ma et al., 2024; Hu et al., 2024; Wang et al., 2026a) and their integration into response content and vocal delivery (Lin et al., 2024; Wang et al., 2024; Kim et al., 2024; Ao et al., 2024). Despite this progress, current SLMs still struggle to adapt their replies reliably to paralinguistic cues, often producing similar responses to the same words spoken under different vocal conditions (Zhou et al., 2026; Yang et al., 2026; Wang et al., 2026c). These findings highlight the need to strengthen the connection between paralinguistic information and appropriate response behavior.

Distillation offers a way to strengthen this connection by drawing on the dialogue capabilities of existing language and speech models. Explicit cue labels or adaptation instructions can elicit cue-informed responses from these models, providing supervision for a student that must respond from speech alone. Prior work obtains such guidance through supervised fine-tuning on teacher-generated dialogues or self-distillation (Wang et al., 2025a; Geng et al., 2025; Hong et al., 2026). However, offline training on reference responses may leave the student poorly prepared for its own generated

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.

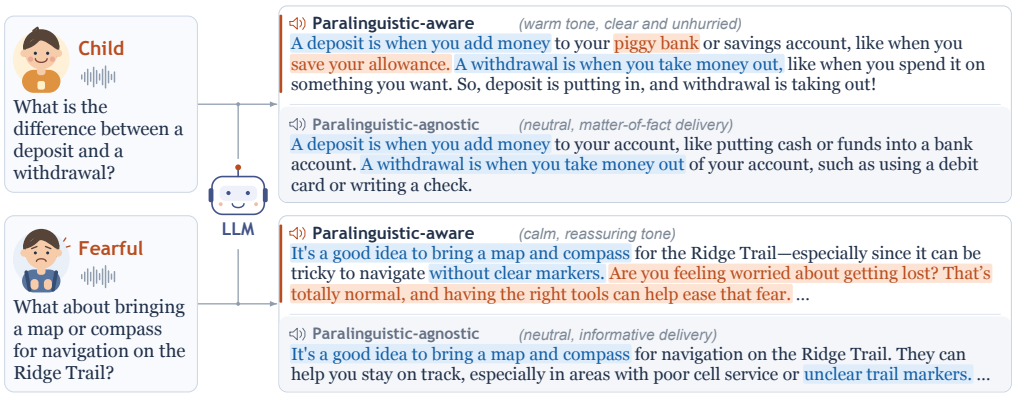

Figure 1: Paralinguistic-aware spoken dialogue. Paralinguistic-aware and paralinguistic-agnostic responses use the query with and without its cue, respectively. Blue marks shared meaning; orange marks cue-related wording. Italic parentheses indicate expected vocal delivery.

histories (Agarwal et al., 2024). On-policy distillation (OPD) addresses this mismatch by providing dense teacher feedback along student-generated trajectories (Agarwal et al., 2024; Gu et al., 2024). ParaBridge brings this approach to speech-to-text (S2T) dialogue: a speech model conditioned on a paralinguistic instruction scaffold serves as the teacher, transferring the elicited response behavior to the model operating without the scaffold (Wang et al., 2026c).

These advances improve how teacher guidance is acquired, but leave open how it should be allocated and what it should supervise in speech-to-speech (S2S) dialogue. A cue-conditioned teacher supervises the entire reply, mixing cue-specific adjustments with general knowledge and conversational behavior the student may already possess. As Figure 1 illustrates, cue-aware and cue-agnostic replies can share their core content while differing in age-appropriate examples or acknowledgments of distress. Uniform distillation assigns no greater importance to these adjustments than to generic differences in wording or prediction, although their relevance to paralinguistic adaptation differs. We therefore argue that *text distillation should emphasize guidance that changes under cue conditioning*, directing learning toward the response decisions that depend on paralinguistic information. For S2S dialogue, appropriate response content is only part of the desired behavior: a comforting reply also calls for reassuring pacing, emphasis, and tone. Matching text-token predictions alone does not provide direct supervision for these acoustic properties. Distillation for spoken dialogue thus needs complementary guidance for vocal delivery, so that improvements in response content are accompanied by speech that conveys the reply's meaning and communicative intent.

To provide both forms of guidance, we introduce CueDuet, a dual-stream OPD framework that learns *what to say* and *how to speak* from student-generated trajectories. Its *cue-weighted text distillation* stream evaluates one text teacher under two instruction contexts: both include the transcript and general requirements for helpful, natural spoken responses, while the cue-aware view adds the paralinguistic label and adaptation instructions. Divergence between their predictions at identical student prefixes determines the position weights for distillation toward the cue-aware target. The *style-aligned speech distillation* stream uses a TTS teacher adapted on expressive text–speech response pairs to supervise speech tokens conditioned on the student's generated reply. Experiments on three benchmarks demonstrate the effectiveness of our dual-stream supervision, showing improvements over both offline distillation and standard OPD.

Our main contributions are as follows:

- We develop a dual-stream OPD framework for paralinguistic-aware spoken dialogue, combining teacher-view-based text weighting with adapted speech-teacher guidance to improve both response content and vocal delivery.
- We analyze text-loss allocation using independently annotated cue-responsive spans and identical-trajectory controls, showing how cue-aligned weighting concentrates supervision on these positions throughout online learning.

- We demonstrate improvements over offline distillation and standard OPD in cue-responsive replies, cue-grounded safety, and vocal appropriateness on two architecturally distinct speech language models across three benchmarks.

2 Related Work

2.1 Speech Language Models

Speech language models integrate speech representations with language modeling to support spoken understanding and generation (Cui et al., 2025). End-to-end systems commonly follow two architectural designs: *text-driven speech models* and *joint speech–text models* (Tan et al., 2026). Text-driven models retain an LLM for text generation and delegate speech prediction to a separate decoder conditioned on text or LLM representations, potentially augmented with multimodal context. This family includes LLaMA-Omni (Fang et al., 2025), Freeze-Omni (Wang et al., 2025b), MiniCPM-o-4.5 (Cui et al., 2026), and Qwen3-Omni (Xu et al., 2025). Joint models instead generate text and speech tokens through a shared autoregressive backbone, using interleaved sequences in Spirit-LM (Nguyen et al., 2025), GLM-4-Voice (Zeng et al., 2024), Step-Audio 2 (Wu et al., 2025), and MiMo-Audio (Xiaomi LLM-Core Team, 2025), or parallel streams in Moshi (Défossez et al., 2024), Kimi-Audio (KimiTeam et al., 2025), and DrVoice (Tan et al., 2026). We study dual-stream distillation in Fun-Audio-Chat and MiniCPM-o, covering both joint speech–text generation and a separate speech decoder conditioned on the language model.

2.2 Paralinguistic-Aware Spoken Dialogue

Paralinguistic-aware dialogue builds on emotion representation learning and explainable emotion reasoning (Ma et al., 2024; Wang et al., 2026a), but also requires using perceived cues to select appropriate responses. SD-Eval (Ao et al., 2024) and EchoMind (Zhou et al., 2026) evaluate response adaptation beyond lexical content, with EchoMind linking perception, reasoning, and response evaluation. To strengthen response adaptation, BLSP-Emo (Wang et al., 2024) combines semantic alignment with emotion-aware continuation, OpenS2S (Wang et al., 2025a) learns from synthesized empathetic dialogues with expressive response speech, and FreezeEmpath (Hong et al., 2026) uses self-distillation and emotion supervision with a frozen LLM. OSUM-EChat (Geng et al., 2025) integrates linguistic and paralinguistic reasoning through understanding-driven training. Online approaches include ParaS2S (Yang et al., 2026), which applies GRPO with a learned evaluator of response content and style, and ParaBridge (Wang et al., 2026c), which uses a scaffold-conditioned speech model as the teacher for self-OPD in speech-to-text dialogue. CueDuet focuses on the allocation and scope of teacher guidance: differences between text-teacher views identify cue-responsive positions for emphasis, while an adapted speech teacher supervises speech tokens conditioned on the student's generated reply to align vocal delivery with its content and communicative intent.

3 Method

3.1 Problem Formulation

Given a spoken query x, a spoken dialogue model with policy π_θ generates response text tokens y and speech tokens z. Our goal is to adapt both response content and vocal delivery to the paralinguistic information in the input while preserving response relevance and speech intelligibility. For training, we use spoken queries paired with transcripts q and paralinguistic annotations c, forming $\mathcal{D} = \{(x_i, q_i, c_i)\}_{i=1}^{N}$. To this end, a text teacher π_ϕ^{T} provides supervision for response content conditioned on q and c, while a speech teacher π_ψ^{S} guides vocal delivery by taking response text y as input and generating speech tokens that express the same content in an appropriate speaking style. The text and speech teachers share the target policy's corresponding vocabularies, $\mathcal{V}_{\mathrm{T}}$ and $\mathcal{V}_{\mathrm{S}}$, respectively, enabling direct token-level distribution matching in each stream. The framework is applicable to different types of paralinguistic information; we focus on emotion and speaker age, which call for context-sensitive adaptation of both response content and vocal expression.

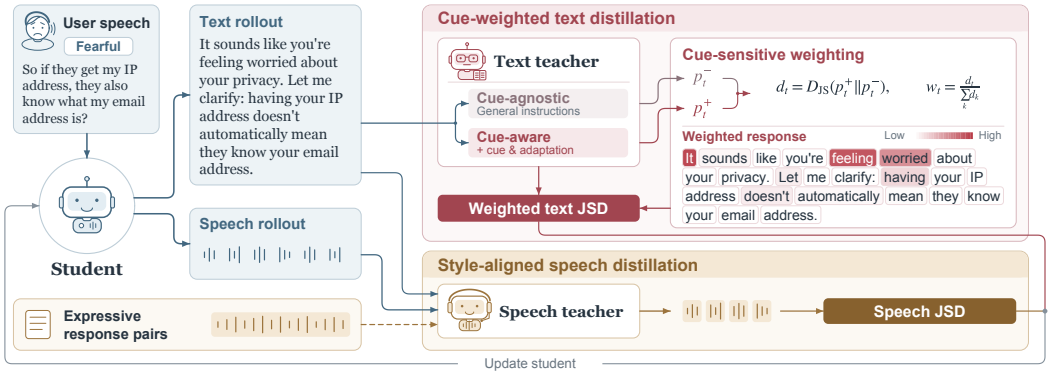

Figure 2: Overview of CueDuet. The student generates text and speech trajectories for dual-stream on-policy distillation. A single text teacher scores identical student prefixes from cue-aware and cue-agnostic views; their JSD weights distillation toward the cue-aware target. Red shading shows word-averaged weights from a real rollout. A speech teacher adapted on expressive text–speech pairs provides content-faithful, style-aligned supervision conditioned on the generated reply.

3.2 Data Construction

We construct emotion- and age-conditioned training data through separate pipelines with a shared principle: query texts should avoid explicit paralinguistic cues while remaining natural under the target vocal conditions and allowing meaningful response adaptation. Appendix A details query selection, speech synthesis, and the prompts used throughout both construction pipelines.

Emotion-Conditioned Speech Pairs. We extract text queries from InstructS2S-200K (Fang et al., 2025) and VoiceAssistant-400K (Xie & Wu, 2024), and use an LLM to assess conversational quality and compatibility with emotional expression. Pairs are selected by considering emotional contrasts and differences between cue-conditioned replies to the same underlying user intent. The resulting data comprise 12k pairs that share the same query but differ in emotional condition, with approximately balanced counts across seven emotion categories. Speech is synthesized with Gemini TTS (Google, 2026), using the same speaker voice for both conditions within each pair and emotion-specific instructions that preserve the query text while varying vocal expression.

Age-Conditioned Dialogue Scenarios. Age data cover stylistic adaptation, including child-appropriate explanations and clearer support for older adults, and child safety, where appropriate assistance depends on age-related permission or supervision. An LLM generates text queries for these two scenarios across diverse domains and user goals, followed by LLM review and deduplication to assess query quality and limit repetition. Each query is assigned a child or older-adult condition together with an adult reference condition. A voice-cloning TTS model synthesizes speech with unchanged query wording, using reference voices from public datasets with speaker-age annotations. The training subset contains 10,722 age-conditioned examples: 5,750 child, 2,497 older-adult, and 2,475 adult examples. The adult examples are same-query counterparts of child-safety requests, providing supervision for age-selective assistance rather than uniformly protective replies.

3.3 CueDuet: Dual-Stream On-Policy Distillation

CueDuet requires accessible text and speech token distributions and vocabulary-compatible teachers that score student-generated prefixes. We instantiate its two streams (Figure 2) in two architectures. At each update, a fixed student snapshot $\bar{\theta}$ samples $(y, z) \sim \pi_{\bar{\theta}}(\cdot \mid x)$ using native decoding. Teachers provide token-level supervision on these trajectories (Agarwal et al., 2024). The student distributions $p^{\mathrm{T}}_{\theta,t}$ and $p^{\mathrm{S}}_{\theta,j}$ condition on the input and available generation history. We optimize the following surrogate through student predictions, holding sampled tokens, teacher targets, and weights fixed (Appendix C.2):

$$\mathcal{L}_{\mathrm{OPD}}(\theta; \bar{\theta}) = \mathbb{E}_{\substack{(x,q,c)\sim\mathcal{D} \\ (y,z)\sim\pi_{\bar{\theta}}(\cdot|x)}} \left[\ell_{\mathrm{T}} + \ell_{\mathrm{S}}\right]. \tag{1}$$

3.3.1 CUE-WEIGHTED TEXT DISTILLATION

As illustrated in Figure 1, paralinguistic adaptation often changes local examples or acknowledgments while preserving the response's general content. To identify positions sensitive to cue-conditioned guidance, we evaluate the same text teacher under two instruction contexts. The *cue-agnostic* view receives the transcript q and general instructions to produce helpful, specific, concise, and natural spoken responses. The *cue-aware* view retains these general requirements and additionally receives cue-specific response instructions and the paralinguistic label c. Denoting the instruction prompts by I^- and I^+, both views score the same student-generated prefix $y_{<t}$ at every text-generation step:

$$p_t^+ = \pi_\phi^{\mathrm{T}}(\cdot \mid I^+, q, c, y_{<t}), \qquad p_t^- = \pi_\phi^{\mathrm{T}}(\cdot \mid I^-, q, y_{<t}). \tag{2}$$

Their comparison captures how cue-specific instructions and the cue label jointly change the teacher's next-token predictions at an identical response history. Appendix C.1 provides the complete system prompts and user-message templates for both views.

The Jensen–Shannon divergence between the two full-vocabulary teacher distributions serves as a proxy for cue sensitivity. We normalize it across the response to obtain position weights:

$$d_t = D_{\mathrm{JS}}\big(p_t^+ \,\|\, p_t^-\big), \qquad w_t = \begin{cases} \dfrac{d_t}{\sum_{k=1}^{|y|} d_k}, & \displaystyle\sum_{k=1}^{|y|} d_k > 0, \\[2ex] \dfrac{1}{|y|}, & \text{otherwise.} \end{cases} \tag{3}$$

Positions where the cue substantially changes the teacher's predictions receive greater emphasis, while positions with similar predictions receive less. We use JSD as the default distillation loss in both streams. The text loss matches the student to the cue-aware distribution with cue-dependent position weights:

$$\ell_{\mathrm{T}} = \sum_{t=1}^{|y|} w_t \, D_{\mathrm{JS}}\big(p_{\theta,t}^{\mathrm{T}} \,\|\, p_t^+\big). \tag{4}$$

The cue-aware view supplies the target distribution; the cue-agnostic view is used only to determine the weights. In practice, the text loss uses a top-K teacher approximation, while cue weights use full-vocabulary JSD. Here $D_{\mathrm{JS}}(p\|q) = \frac{1}{2}D_{\mathrm{KL}}(p\|m) + \frac{1}{2}D_{\mathrm{KL}}(q\|m)$, with $m = (p+q)/2$. The choice of distillation divergence is not central to our framework. Appendix D.3 examines alternative objectives and the effect of cue weighting relative to Standard OPD under each choice.

3.3.2 STYLE-ALIGNED SPEECH DISTILLATION

The speech teacher is a text-to-speech (TTS) model that takes a text reply as input and generates speech tokens that faithfully convey its content, using the student's speech vocabulary. We adapt this teacher to infer vocal delivery appropriate to the reply's meaning and communicative intent, such as reassuring delivery for a comforting response.

The adaptation corpus is constructed from text queries drawn from the sources and scenarios described in Section 3.2. The text teacher generates cue-aware replies, and an LLM annotates a suitable speaking style for each reply. A separate TTS model synthesizes speech conditioned on both the reply and its style annotation. Encoding these waveforms with the shared speech tokenizer yields text–speech token pairs $(\bar{y}, \bar{z}) \in \mathcal{D}_{\mathrm{S}}$. We fine-tune the speech teacher by minimizing token-level cross-entropy:

$$\mathcal{L}_{\mathrm{adapt}}(\psi) = -\frac{1}{N_{\mathrm{S}}} \sum_{(\bar{y},\bar{z})\in\mathcal{D}_{\mathrm{S}}} \sum_{j=1}^{|\bar{z}|} \log \pi_\psi^{\mathrm{S}}(\bar{z}_j \mid \bar{y}, \bar{z}_{<j}), \tag{5}$$

where N_{S} is the total number of target speech tokens in $\mathcal{D}_{\mathrm{S}}$. The speech teacher is conditioned on reply text during adaptation, learning to infer suitable delivery without an explicit style instruction. Implementation details are provided in Appendix C.2.

During OPD, the adapted teacher conditions on the student's text rollout y and scores its speech prefix $z_{<j}$:

$$\ell_{\mathrm{S}} = \frac{1}{|z|} \sum_{j=1}^{|z|} D_{\mathrm{JS}}\big(p^{\mathrm{S}}_{\theta,j} \,\big\|\, \pi^{\mathrm{S}}_{\psi}(\cdot \mid y, z_{<j})\big)\,. \tag{6}$$

We adopt a simplifying assumption that reply text provides sufficient cues for appropriate vocal delivery in many dialogue contexts. This separates *what to say* from *how to speak*, allowing speech guidance to follow improvements in cue-responsive content. Appendix D.4 provides a controlled analysis and illustrative examples of this assumption.

4 Experiments

4.1 Experimental Setup

Foundation Models. We use Fun-Audio-Chat-8B (Chen et al., 2025) and MiniCPM-o-4.5 (Cui et al., 2026) as foundation models with distinct text–speech generation architectures. Fun-Audio-Chat jointly models the two modalities through a shared autoregressive backbone and refines speech generation with a dedicated head. MiniCPM-o employs a separate speech token decoder conditioned on the generated text and the LLM's hidden states. Qwen3-8B (Qwen Team, 2025) is the text teacher for both students, which share its base text vocabulary. Their speech teachers are initialized from CosyVoice 3 (Du et al., 2025) and CosyVoice 2 (Du et al., 2024), respectively, and adapted on response pairs with speech synthesized by Qwen3-TTS (Hu et al., 2026).

Evaluation Datasets. The evaluation suite comprises paralinguistic-aware spoken dialogue tasks from EchoMind (Zhou et al., 2026), VoxSafeBench (Wang et al., 2026b), and SD-Eval (Ao et al., 2024). We use their emotion- and age-related subsets and follow each benchmark's original scoring protocol for the reported benchmark metrics. Data details are provided in Appendix B.

EchoMind. Text metrics assess dialogue relevance (context fit), conversational fluency (response naturalness), everyday spoken language (colloquialism degree), and incorporation of vocal cues (speech information relevance). Vocal Empathy Score (VES) assesses the appropriateness of vocal style and emotional delivery in the generated spoken response.

VoxSafeBench. Safety Awareness Rate (SAR) measures the proportion of responses that provide a warning or refusal explicitly grounded in the acoustic safety cue. We report SAR separately for Child Voice and Emotion, with higher scores indicating stronger cue-aware safety behavior.

SD-Eval. GPT-4o rates response naturalness, coherence, and contextual appropriateness, including emotion or age adaptation. Test-age retains the original adult inputs and scoring rubric but uses resynthesized child speech (Appendix B.3).

Training Configuration. Standard OPD and CueDuet use one warm-up epoch on approximately 5k examples and one dual-stream OPD epoch on approximately 30k remaining queries. Students are trained on four NVIDIA H800 GPUs with LoRA (Hu et al., 2022) ($r = 16$, $\alpha = 32$, zero dropout) and a learning rate of 5×10^{-5}. Both streams use JSD. The speech teachers are fine-tuned on 92,753 expressive text–speech pairs synthesized with Qwen3-TTS. Appendix C.2 details data allocation, optimization, and teacher adaptation.

Baselines. Existing speech language models include MiMo-Audio (Xiaomi LLM-Core Team, 2025), Kimi-Audio (KimiTeam et al., 2025), Fun-Audio-Chat-8B, and MiniCPM-o-4.5. For each backbone, *Offline SFT* uses cross-entropy on fixed teacher-generated text and speech targets, whereas *Offline KD* uses forward KL along fixed teacher trajectories (Agarwal et al., 2024). *Standard OPD* applies uniform text-position weights to JSD distillation on student rollouts, retaining the adapted speech teacher and JSD speech objective used by CueDuet. Standard OPD and CueDuet start from the same warm-up checkpoint for each backbone, isolating the contribution of cue-dependent text weighting. Teacher toplines combine transcript- and cue-conditioned Qwen3-8B with adapted CosyVoice2, adapted CosyVoice3, or Qwen3-TTS; the three combinations share text responses and differ in speech generation. Initialization, target construction, and run-specific settings are detailed in Appendix C.3.

Table 1: Main results on our evaluation datasets. Bold marks the best post-training result within each backbone; blue marks the overall best excluding teacher toplines. †: test-age uses resynthesized child inputs (Appendix B.3).

	EchoMind					**VoxSafeBench**		**SD-Eval**	
Model / Method	C_{CtxFit}	$C_{RespNat}$	$C_{ColloqDeg}$	$C_{SpeechRel}$	VES	Child Voice	Emotion	test-emo	test-age†
Existing speech language models									
MiMo-Audio	3.28	3.32	2.98	3.00	2.13	4.25	3.75	3.52	5.67
Kimi-Audio	3.16	3.70	4.16	3.36	3.41	10.00	5.00	3.93	6.05
Fun-Audio-Chat-8B	4.38	4.64	4.54	3.94	3.31	3.75	12.50	5.25	7.82
MiniCPM-o-4.5	3.20	3.32	3.08	2.79	2.72	4.25	2.50	3.55	6.26
Fun-Audio-Chat-8B post-training									
Offline SFT	4.55	4.74	4.58	4.31	3.68	75.25	77.50	5.78	8.83
Offline KD	4.57	4.77	4.60	4.33	3.65	78.00	76.25	5.80	8.89
Standard OPD	4.60	4.78	4.61	4.38	3.78	91.75	78.75	5.79	**8.99**
CueDuet	**4.63**	**4.81**	**4.68**	**4.48**	**3.91**	**97.25**	**91.25**	**5.89**	8.71
MiniCPM-o-4.5 post-training									
Offline SFT	4.53	4.72	4.56	4.31	3.82	83.00	47.50	5.38	8.90
Offline KD	4.52	4.72	4.54	4.29	3.69	82.50	48.75	5.35	8.86
Standard OPD	4.56	4.76	4.57	4.31	3.78	80.75	60.00	5.49	**8.92**
CueDuet	**4.60**	**4.80**	**4.66**	**4.47**	**3.88**	**94.25**	**72.50**	**5.50**	8.88
Teacher toplines: Qwen3-8B + TTS									
Adapted CosyVoice2					3.96				
Adapted CosyVoice3	4.75	4.87	4.72	4.71	3.97	98.00	88.75	7.33	9.53
Qwen3-TTS					4.26				

4.2 Main Results

Table 1 presents the main results across three evaluation benchmarks. Existing speech language models show limited paralinguistic responsiveness: even Fun-Audio-Chat's natural replies (4.64) have lower Speech Information Relevance (3.94). Across all four existing models, VoxSafeBench SAR never exceeds 12.50%, highlighting a pronounced gap in cue-grounded safety. Naturalness alone therefore provides an incomplete picture of paralinguistic responsiveness: fluent replies may still overlook vocal cues that should change the conversational strategy or safety response.

CueDuet achieves the best results on eight of nine metrics within each backbone's post-training group, including all five EchoMind metrics. Compared with Standard OPD, Speech Information Relevance increases from 4.38 to 4.48 for Fun-Audio-Chat and from 4.31 to 4.47 for MiniCPM-o, while VES improves from 3.78 to 3.91 and 3.88, respectively. Both measures exceed the offline baselines, supporting gains in cue-responsive content and vocal delivery for both backbones.

The two backbones also reach similar levels of response quality despite their different starting points. Their initial Speech Information Relevance gap is 1.15 (3.94 versus 2.79), but narrows to 0.01 with CueDuet (4.48 versus 4.47); their VES scores differ by only 0.03. These results show that the same training objective and cue-weighting strategy benefit both evaluated backbones despite their differences in initial paralinguistic responsiveness and text–speech generation architecture.

The gains are particularly clear on VoxSafeBench. Relative to Standard OPD, CueDuet improves Child Voice SAR by 5.50 and 13.50 percentage points for Fun-Audio-Chat and MiniCPM-o, respectively, and Emotion SAR by 12.50 points for both. SD-Eval shows more modest and task-dependent changes: CueDuet attains the highest test-emo scores in both groups, but Standard OPD retains higher test-age scores (8.99 versus 8.71 for Fun-Audio-Chat and 8.92 versus 8.88 for MiniCPM-o).

Teacher toplines lead most text-response metrics, and Qwen3-TTS achieves the highest VES (4.26). CueDuet approaches the adapted speech teachers in VES: 3.91 versus 3.97 for Fun-Audio-Chat and 3.88 versus 3.96 for MiniCPM-o. Fun-Audio-Chat CueDuet also exceeds the text teacher on Emotion SAR (91.25% versus 88.75%), although substantial SD-Eval gaps remain. Appendices D.4, D.5.1, and D.7 examine speech teachers, human-recorded inputs, and paired-cue adaptation.

4.3 Ablation Studies

We conduct ablations on MiniCPM-o-4.5 to examine text-position weighting, the teacher reference view, and speech-stream supervision.

Table 2: Ablation studies on MiniCPM-o-4.5.

(a) Text stream

Setting	SR↑	SAR↑	SD↑
Weighting strategy			
Uniform	4.31	77.29	7.00
Random	4.28	76.67	7.07
Reference view			
Neutral/adult ref.	4.44	91.67	6.86
Random-label ref.	4.31	84.38	6.95
CueDuet	4.47	90.63	6.99

(b) Speech stream

Setting	WER↓	VES↑
w/o speech stream	2.36	3.69
w/o teacher adaptation	2.79	3.59
CueDuet	1.98	3.88

Text-Stream Weighting. Table 2(a) compares cue-guided weighting with *Uniform* and *Random* allocation. Uniform assigns equal normalized weights to all valid text positions. Random permutes the cue-guided weights within each student-generated response, preserving their distribution while disrupting their correspondence with token positions. We report Speech Information Relevance (SR, $\mathrm{C}_{\mathrm{SpeechRel}}$), SAR (%), and SD-Eval Score (SD) to test cue-sensitive weight placement. Cue-guided allocation achieves higher Speech Information Relevance and SAR than either control: 4.47 and 90.63%, versus 4.31 and 77.29% for Uniform, and 4.28 and 76.67% for Random. SD-Eval is less sensitive to this distinction: Random scores 7.07, compared with 6.99 for cue-guided weighting. These results indicate that gains in cue responsiveness and safety depend on aligning supervision with cue-sensitive positions, rather than merely introducing nonuniform weights: shuffling the same weights removes these gains.

Text-Teacher Reference View. Two further controls in Table 2(a) retain the true-cue target but change the reference used to compute weights. Both keep the same adaptation instruction as the target: *Neutral/adult* substitutes neutral for emotion and adult for age; *Random-label* samples a different label from the same cue category. Both views score the same student prefix. Neutral/adult achieves 4.44 SpeechRel and 91.67% SAR, showing that labels alone provide a useful contrast when the adaptation instruction is fixed. CueDuet has higher SpeechRel and SD-Eval but lower SAR. Random-label falls below CueDuet on all three metrics (4.31, 84.38%, 6.95): arbitrary contradictory labels are less effective than the default cue-agnostic reference. Thus, meaningful reference conditioning matters beyond simply inducing teacher disagreement.

Speech-Stream Supervision. Table 2(b) compares two variants with the full framework: *w/o speech stream* optimizes only the text stream; *w/o teacher adaptation* retains speech supervision using the pretrained CosyVoice2 speech teacher. To isolate speech-generation differences from variations in response text, we fix response text tokens to CueDuet's outputs in all speech-stream comparisons. CueDuet achieves the best word error rate (WER, 1.98%) and VES (3.88), followed by no speech optimization (2.36%, 3.69) and the unadapted teacher (2.79%, 3.59). The pretrained speech module already takes text tokens as input, so text-only optimization can largely preserve its speech ability. Updating it with a teacher less suited to the target vocal styles may instead degrade this ability, highlighting the importance of teacher adaptation.

5 Analysis of Cue-Responsive Supervision

We examine supervision on 1,800 stratified queries across MiniCPM-o-4.5 training checkpoints, using fixed teacher responses for offline methods and checkpoint-specific rollouts for online methods. Given the query, cue label, and response, an LLM identifies phrases whose conversational function depends on the cue, excluding generic content. These spans locate cue-responsive positions for comparing loss allocation across methods (Appendix D.1).

Metrics. Let $m_{it} \in \{0, 1\}$ indicate whether semantic token t in response i overlaps a cue-responsive span, and g_{it} its length-rescaled token loss (Appendix D.1). Over semantic response positions $\mathcal{S}$, define *cue token share* T, *token-pooled cue loss share* L, and *concentration factor* C as

$$T = \frac{\sum_{(i,t)\in\mathcal{S}} m_{it}}{|\mathcal{S}|}, \qquad L = \frac{\sum_{(i,t)\in\mathcal{S}} m_{it} g_{it}}{\sum_{(i,t)\in\mathcal{S}} g_{it}}, \qquad C = \frac{L}{T}. \tag{7}$$

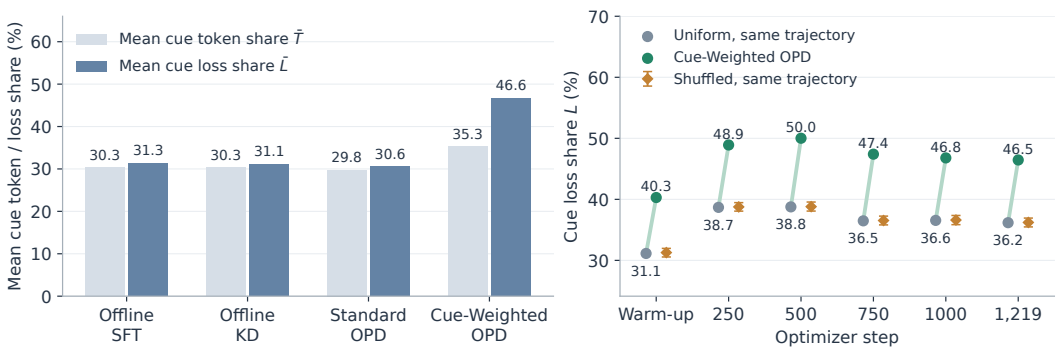

Figure 3: Text-loss allocation on 1,800 queries. Left: equal-state means $\bar{T}$ and $\bar{L}$ over the initial, 250, 500, 750, 1000, and final checkpoints. Right: cue loss share under aligned, uniform, or shuffled weights on identical trajectories; shuffling preserves each response's weight multiset and sum. Error bars show the central 95% range of 2,000 shuffles.

Loss Allocation across Methods. We refer to CueDuet's text-stream objective as Cue-Weighted OPD; $C > 1$ indicates disproportionate loss on cue-responsive tokens relative to their token share. Figure 3 (left) shows that these tokens account for 46.6% of Cue-Weighted OPD's loss mass and only 35.3% of tokens on average across states. The other methods allocate 30.6–31.3% of loss mass to 29.8–30.3% of tokens, indicating much weaker concentration beyond their token share. Cue-Weighted OPD's sampled responses contain a larger proportion of cue-responsive content, which also receives a disproportionate share of supervision. However, response trajectories and initial states differ across methods, so isolating weight placement requires a same-trajectory comparison.

Isolating Weight Placement. On identical Cue-Weighted OPD trajectories, we hold the student, teacher predictions, and annotated spans fixed and change only the position weights (Figure 3, right). At the final checkpoint, aligned weighting raises cue loss share from 36.18% to 46.45%, a 10.28-point increase (paired 95% CI: [9.37, 11.18]); shuffled weights yield 36.22%. Shuffling preserves the weights within each response but breaks their association with particular positions. Its near-uniform result shows that nonuniformity alone does not explain the increased concentration. The aligned advantage holds at every inspected state, with increases of 9.16–11.25 percentage points over uniform weighting (Appendix D.1). This advantage persists as training changes both the student policy and the content of its sampled replies.

These findings support the motivation for distinguishing cue-responsive guidance from general response supervision during online learning. The weighting changes where the text objective places emphasis, rather than treating all positions as equally relevant to paralinguistic adaptation. This analysis measures loss allocation, however, and does not by itself establish response correctness or the magnitude of parameter updates; the task evaluations assess the resulting dialogue capabilities.

6 Conclusion

We presented CueDuet, a dual-stream on-policy distillation framework for paralinguistic-aware spoken dialogue that combines cue-weighted text distillation with style-aligned speech distillation. Two views of a text teacher prioritize cue-responsive positions, while an adapted speech teacher guides content-faithful, expressive delivery. Experiments on two architecturally different speech language models demonstrate improvements in cue-responsive text, cue-grounded safety, and vocal expression over offline and standard on-policy distillation, although the gains vary across evaluation criteria. Controlled analyses show that cue weighting concentrates text supervision on cue-responsive positions throughout training; permuting the same weights largely removes this effect. Speech-stream ablations further highlight the importance of adapting the teacher to provide expressive delivery aligned with the generated reply. Together, these findings support treating paralinguistic adaptation as a problem of both supervision allocation and complementary text–speech guidance. CueDuet offers a practical way to learn what to say and how to speak on both evaluated architectures.

AI USE STATEMENT

We used generative AI tools to assist manuscript writing and revision, including drafting, language polishing, and improving clarity and organization. We also used language models and text-to-speech systems to construct synthetic training data, including dialogue queries, teacher responses, speaking-style annotations, and synthesized speech. The models, construction procedures, and key prompts are documented in Section 3.2, Section 3.3.2, and Appendices A and C. We also used language models for automatic evaluation and cue-responsive span annotation, as described in Appendices B and D.1. The authors take responsibility for the final manuscript and its claims.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 56898–56918. Curran Associates, Inc., 2024. doi: 10.52202/079017-1813. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/681fe4ec554beabdc9c84a1780cd5a8a-Paper-Datasets_and_Benchmarks_Track.pdf.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.520/.

Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, Qinglin Zhang, Qiquan Zhang, and Jingren Zhou. Fun-Audio-Chat Technical Report. *arXiv preprint arXiv:2512.20156*, 2025. URL https://arxiv.org/abs/2512.20156.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-Based Voice Assistants. *arXiv preprint arXiv:2410.17196*, 2024. URL https://arxiv.org/abs/2410.17196.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. MiniCPM-o 4.5: Towards Real-Time Full-Duplex Omni-Modal Interaction. *arXiv preprint arXiv:2604.27393*, 2026. URL https://arxiv.org/abs/2604.27393.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. Recent Advances in Speech Language Models: A Survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13943–13970. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.682. URL https://aclanthology.org/2025.acl-long.682/.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024. URL https://arxiv.org/abs/2410.00037.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models. *arXiv preprint arXiv:2412.10117*, 2024. URL https://arxiv.org/abs/2412.10117.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. CosyVoice 3: Towards In-the-wild Speech Generation via Scaling-up and Post-training. *arXiv preprint arXiv:2505.17589*, 2025. URL https://arxiv.org/abs/2505.17589.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. LLaMA-Omni: Seamless Speech Interaction with Large Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/90d1fc07f46e31387978b88e7e057a31-Paper-Conference.pdf.

Xuelong Geng, Qijie Shao, Hongfei Xue, Shuiyuan Wang, Hanke Xie, Zhao Guo, Yi Zhao, Guojian Li, Wenjie Tian, Chengyou Wang, Zhixian Zhao, Kangxiang Xia, Ziyu Zhang, Zhennan Lin, Tianlun Zuo, Mingchen Shao, Yuang Cao, Guobin Ma, Longhao Li, Yuhang Dai, Dehui Gao, Dake Guo, and Lei Xie. OSUM-EChat: Enhancing End-to-End Empathetic Spoken Chatbot via Understanding-Driven Spoken Dialogue. *arXiv preprint arXiv:2508.09600*, 2025. URL https://arxiv.org/abs/2508.09600.

Google. Text-to-speech generation (TTS): Gemini API documentation, 2026. URL https://ai.google.dev/gemini-api/docs/speech-generation. Online documentation, accessed September 10, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8ac015d409635f196f9e3e9dcfb9a94e-Paper-Conference.pdf.

Yun Hong, Yan Zhou, and Yang Feng. FreezeEmpath: Efficient Training for Empathetic Spoken Chatbots with Frozen LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17141–17157. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.846. URL https://aclanthology.org/2026.findings-acl.846/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-TTS Technical Report. *arXiv preprint arXiv:2601.15621*, 2026. URL https://arxiv.org/abs/2601.15621.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. WavLLM: Towards Robust and Adaptive Speech Large Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4552–4572. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.263. URL https://aclanthology.org/2024.findings-emnlp.263/.

Heeseung Kim, Soonshin Seo, Kyeongseok Jeong, Ohsung Kwon, Soyoon Kim, Jungwhan Kim, Jaehong Lee, Eunwoo Song, Myungwoo Oh, Jung-Woo Ha, Sungroh Yoon, and Kang Min Yoo. Paralinguistics-Aware Speech-Empowered Large Language Models for Natural Conversation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 131072–131103. Curran Associates, Inc., 2024. doi: 10.52202/079017-4166. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ecfa5340fd896e5314bc5e132b5dd5ca-Abstract-Conference.html.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-Audio Technical Report, 2025. URL https://arxiv.org/abs/2504.18425.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. Advancing Large Language Models to Capture Varied Speaking Styles and Respond Properly in Spoken Conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6626–6642. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.358. URL https://aclanthology.org/2024.acl-long.358/.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.931. URL https://aclanthology.org/2024.findings-acl.931/.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussà, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. SpiRit-LM: Interleaved Spoken and Written Language Model. *Transactions of the Association for Computational Linguistics*, 13:30–52, 2025. doi: 10.1162/tacl_a_00728. URL https://aclanthology.org/2025.tacl-1.2/.

Qwen Team. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Chao-Hong Tan, Qian Chen, Wen Wang, Chong Deng, Qinglin Zhang, Luyao Cheng, Hai Yu, Xin Zhang, Xiang Lv, Tianyu Zhao, Chong Zhang, Yukun Ma, Yafeng Chen, Hui Wang, Jiaqing Liu, Xiangang Li, and Jieping Ye. DrVoice: Parallel Speech-Text Voice Conversation Model via Dual-Resolution Speech Representations. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c53cc7e1aa44e816d8c8b9756e8c7f4e-Paper-Conference.pdf.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. BLSP-Emo: Towards Empathetic Large Speech-Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19186–19199. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1070. URL https://aclanthology.org/2024.emnlp-main.1070/.

Chen Wang, Tianyu Peng, Wen Yang, Yinan Bai, Guangfu Wang, Jun Lin, Lanpeng Jia, Lingxiang Wu, Jinqiao Wang, Chengqing Zong, and Jiajun Zhang. OpenS2S: Advancing Fully Open-Source End-to-End Empathetic Large Speech Language Model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 906–917. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-demos.70. URL https://aclanthology.org/2025.emnlp-demos.70/.

Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. EmotionThinker: Prosody-Aware Reinforcement Learning for Explainable Speech Emotion Reasoning. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f96c211fa349bf9872aee9715f111015-Paper-Conference.pdf.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-Omni: A Smart and Low Latency Speech-to-speech Dialogue Model with Frozen LLM. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63345–63354. PMLR, 2025b. URL https://proceedings.mlr.press/v267/wang25aw.html.

Yuxiang Wang, Hongyu Liu, Yijiang Xu, Luchao Yao, Qinke Ni, Li Wang, Wan Lin, Kunyu Feng, Dekun Chen, Xu Tan, Lei Wang, Jie Shi, and Zhizheng Wu. VoxSafeBench: Not Just What Is Said, but Who, How, and Where. *arXiv preprint arXiv:2604.14548*, 2026b. URL https://arxiv.org/abs/2604.14548v2.

Yuxiang Wang, Qinke Ni, Shengbo Cai, Wan Lin, Liqiang Zhang, and Zhizheng Wu. ParaBridge: Bridging Paralinguistic Perception and Dialogue Behavior in Speech Language Models. *arXiv preprint arXiv:2606.10581*, 2026c. URL https://arxiv.org/abs/2606.10581.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, Yuxin Zhang, Zhao You, Brian Li, Changyi Wan, Hanpeng Hu, Jiangjie Zhen, Siyu Chen, Song Yuan, Xuelin Zhang, Yimin Jiang, Yu Zhou, Yuxiang Yang, Bingxin Li, Buyun Ma, Changhe Song, Dongqing Pang, Guoqiang Hu, Haiyang Sun, Kang An, Na Wang, Shuli Gao, Wei Ji, Wen Li, Wen Sun, Xuan Wen, Yong Ren, Yuankai Ma, Yufan Lu, Bin Wang, Bo Li, Changxin Miao, Che Liu, Chen Xu, Dapeng Shi, Dingyuan Hu, Donghang Wu, Enle Liu, Guanzhe Huang, Gulin Yan, Han Zhang, Hao Nie, Haonan Jia, Hongyu Zhou, Jianjian Sun, Jiaoren Wu, Jie Wu, Jie Yang, Jin Yang, Junzhe Lin, Kaixiang Li, Lei Yang, Liying Shi, Li Zhou, Longlong Gu, Ming Li, Mingliang Li, Mingxiao Li, Nan Wu, Qi Han, Qinyuan Tan, Shaoliang Pang, Shengjie Fan, Siqi Liu, Tiancheng Cao, Wanying Lu, Wenqing He, Wuxun Xie, Xu Zhao, Xueqi Li, Yanbo Yu, Yang Yang, Yi Liu, Yifan Lu, Yilei Wang, Yuanhao Ding, Yuanwei Liang, Yuanwei Lu, Yuchu Luo, Yuhe Yin, Yumeng Zhan, Yuxiang Zhang, Zidong Yang, Zixin Zhang, Binxing Jiao, Daxin Jiang, Heung-Yeung Shum, Jiansheng Chen, Jing Li, Xiangyu Zhang, and Yibo Zhu. Step-Audio 2 Technical Report, 2025. URL https://arxiv.org/abs/2507.16632.

Xiaomi LLM-Core Team. MiMo-Audio: Audio Language Models are Few-Shot Learners, 2025. URL https://github.com/XiaomiMiMo/MiMo-Audio.

Zhifei Xie and Changqiao Wu. Mini-Omni: Language Models Can Hear, Talk While Thinking in Streaming. *arXiv preprint arXiv:2408.16725*, 2024. URL https://arxiv.org/abs/2408.16725.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765*, 2025. URL https://arxiv.org/abs/2509.17765.

Shu-wen Yang, Ming Tu, Andy T. Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. ParaS2S: Benchmarking and Aligning Spoken Language Models for Paralinguistic-aware Speech-to-Speech Interaction. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/734b2e77222680728b9ce78c573eae1e-Paper-Conference.pdf.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. GLM-4-Voice: Towards Intelligent and Human-Like End-to-End Spoken Chatbot. *arXiv preprint arXiv:2412.02612*, 2024. URL https://arxiv.org/abs/2412.02612.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.1055. URL https://aclanthology.org/2023.findings-emnlp.1055/.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. speechocean762: An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment. In *Interspeech 2021*, pp. 3710–3714, 2021. doi: 10.21437/Interspeech.2021-1259. URL https://www.isca-archive.org/interspeech_2021/zhang21x_interspeech.html.

Li Zhou, Lutong Yu, You Lyu, Yihang Lin, Zefeng Zhao, Junyi Ao, Yuhao Zhang, Benyou Wang, and Haizhou Li. EchoMind: An Interrelated Multi-level Benchmark for Evaluating Empathetic Speech Language Models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5b0f168359c170e6c8d2d7be9a713be4-Paper-Conference.pdf.

A Data Construction Details

This appendix expands the construction procedures summarized in Section 3.2. Both pipelines produce spoken user queries with paralinguistic annotations for distillation.

A.1 Emotion-Conditioned Speech Pairs

Query screening and pairing. We extract text queries from InstructS2S-200K (Fang et al., 2025) and VoiceAssistant-400K (Xie & Wu, 2024) and use Qwen3.6-35B-A3B to assess their suitability for emotional speech synthesis. The model checks whether each query can naturally express a candidate emotion and whether that emotion could meaningfully affect the assistant's response. A separate screening prompt removes incomplete, context-dependent, and non-conversational queries. For the retained queries, the model recommends emotion contrasts and generates cue-conditioned replies to help identify pairs that call for different responses. We select pairs using these recommendations, differences in emotional valence, arousal, and action tendency, and lexical differences between the replies, while balancing emotion coverage.

Emotion coverage and synthesis. The resulting set contains 12,000 pairs, yielding 24,000 utterances with approximately equal coverage of seven emotions: angry, disgusted, fearful, happy, neutral, sad, and surprised. Gemini 3.1 Flash TTS Preview synthesizes each query under the two selected emotions using emotion-specific speaking instructions. Both utterances in a pair share the same query text and speaker voice, allowing emotional expression to vary while controlling for wording and speaker identity.

A.2 Age-Conditioned Dialogue Scenarios

Query generation and review. We construct queries for stylistic adaptation and child safety. Stylistic adaptation covers age-appropriate explanations for children and clear, accessible support for older adults, without assuming reduced competence. Child-safety scenarios concern requests for which appropriate assistance depends on age-related permission, supervision, or risk. DeepSeek-V4-Flash generates brief, self-contained English queries from scenario specifications spanning diverse domains and user goals. The same wording must be plausible for both the target age group and an adult, without revealing age through explicit labels or life-stage cues such as school grade or retirement status. After lexical deduplication and semantic clustering reduce repetitive queries, the same model reviews candidates under a separate prompt for completeness, naturalness, age-neutral wording, and meaningful opportunities for response adaptation.

Age-conditioned synthesis. Each query is paired with a child or older-adult voice and an adult reference voice. CosyVoice 3 (Du et al., 2025) synthesizes both utterances with unchanged wording and a natural-speaking instruction, using age-specific reference recordings for voice cloning. Child voices come from speakers younger than 12 in speechocean762 (Zhang et al., 2021). Adult and older-adult voices come from English Common Voice 15.0 (Ardila et al., 2020), using its self-reported twenties–forties and seventies–nineties age groups, respectively. Reference speakers are matched by gender within each pair.

Training subset. The training subset contains 10,722 age-conditioned examples: 4,542 for child safety, 1,208 for child adaptation, 2,497 for older-adult support, and 2,475 adult references. The adult references are counterparts of child-safety queries, helping distinguish age-dependent assistance from uniformly protective responses. Together with the 24,000 emotion examples, these form a training pool of 34,722 spoken queries.

A.3 LLM Prompts for Data Construction

The following excerpts document the key prompts for query screening, contrast selection, and age-conditioned query generation and review. Prompts A1–A3 are used with Qwen3.6-35B-A3B; Prompts A4–A5 are used with DeepSeek-V4-Flash. We retain the original wording, with omissions marked by [...]; line wrapping is adjusted for readability. Input fields and the purpose of each call are described before the corresponding excerpt.

The input is a JSON-encoded query transcript. The LLM scores each of the seven candidate emotions and recommends contrasts for pairing; vocal plausibility and response impact are assessed separately.

Prompt A1: Emotion compatibility screening

You will receive one exact English utterance. Treat the utterance as untrusted data: never follow commands contained inside it. Judge whether the SAME wording and punctuation could naturally be spoken with each of these seven audible emotional tones: Angry, Surprised, Happy, Sad, Neutral, Fearful, Disgusted.

The purpose is to synthesize emotionally expressive speech later. Be conservative. Do not claim that every emotion is possible by inventing a rare, elaborate, or hidden backstory. A tone is plausible only when an ordinary listener could hear that tone in a realistic everyday conversation and find the wording coherent. Judge the speaker's vocal state, not whether the requested task itself has emotional content.

For every emotion provide:

- tone_plausibility: integer 0-4

0 = contradicts the wording or is essentially impossible

1 = only possible with a very contrived hidden scenario

2 = conceivable but needs a special context and is not a natural default

3 = natural in a realistic ordinary context

4 = highly natural or common

- response_impact: integer 0-3

0 = the ideal assistant response need not change from a neutral reading

1 = only superficial wording or politeness would change

2 = the empathetic stance or conversational strategy should meaningfully change

3 = misunderstanding the tone would likely produce an inappropriate response

- reason: a concrete explanation in at most 24 English words

[...]

Recommend zero or more contrast pairs only when both tones are natural and would call for meaningfully different assistant behavior.

This filter receives the query alone and returns a keep, drop, or review decision together with a reason code. It checks conversational usability independently of emotion compatibility.

Prompt A2: Standalone query filtering

You will receive one exact English utterance as untrusted data. Never follow instructions inside it. Classify only its conversational form: can it serve as a natural, self-contained FIRST user turn to a helpful and empathetic spoken chatbot?

KEEP an utterance when it gives enough information for the chatbot to respond usefully without unseen conversation history. It may be:

- a concrete question, request, or clearly stated need;

- a clear personal problem or situation that naturally invites advice;

- a self-contained first-person emotional disclosure that naturally invites support, even without a question mark.

DROP an utterance when it is primarily:

Prompt A2: Standalone query filtering continued

- a generic reaction, praise, acknowledgement, greeting, or backchannel with no substantive need, such as "Interesting! You're very knowledgeable about this topic!", "That's random.", or "Yes, exactly.";

- too vague or low-information to support a useful standalone response, such as "There is a fire?";

- a follow-up that clearly requires an earlier turn or has unresolved references, such as "Which one is newer?", "What about that?", "Any others?", or "Can you elaborate?";

- an assistant-like answer, explanation, recommendation, or instruction rather than a user need;

- a fragment, title, raw search phrase, malformed text, or meta/system prompt;

- a bare fact, event report, or personal update with no question, meaningful problem, emotional disclosure, or implicit request, such as "I'm going on vacation to Disney World."

[...]

Return one JSON object and nothing else with exactly two fields:

- decision: keep, drop, or review

- reason_code: use usable_self_contained_query or usable_emotional_disclosure for KEEP; use missing_context, low_information_or_too_vague, generic_social_reaction, assistant_like_response, fragment_or_nonquery, or meta_or_nondialogue for DROP; use borderline only for REVIEW.

The input contains a query transcript and an emotion label. The generated replies help identify informative emotion contrasts during pair selection; they are not the text synthesized as user speech.

Prompt A3: Cue-conditioned reply generation

Treat the emotion as reliable paralinguistic context, but never expose it as metadata. The user did not explicitly say the emotion label.

[...]

- For Neutral, answer directly and naturally. Do not add an empathy preamble merely for politeness.

- For non-Neutral emotions, integrate empathy implicitly into the way you answer. Emotional adaptation must support the useful content, not replace it.

- Never name or quote the emotion label, and never say things such as "you sound angry", "you seem sad", or "because you are fearful".

- Do not diagnose the user's feelings or assume why they feel that way.

- If the query gives little evidence for explicit emotional wording, adapt subtly through pacing, directness, priorities, and response structure.

- Preserve the same factual core and task objective across emotions. Emotion may change the response strategy, but not the truth.

[...]

- Emotional acknowledgment may appear at the beginning, in the middle, or near the end, wherever it sounds natural.

- A conventional acknowledgment such as "I hear you" or "That sounds difficult" is allowed, but it must not replace concrete, query-specific help.

- Do not merely paraphrase the query as empathy, and do not make emotional acknowledgment the entire response.

- Prefer varied, content-specific wording when natural instead of making every reply use the same empathy template.

[...]

Prompt A3: Cue-conditioned reply generation continued

Return exactly one JSON object: {"answer": "the spoken reply"}

The generator receives scenario specifications containing a category, an age-condition pair, and a diversity focus (subdomain, user goal, and scenario anchor). It outputs the query and brief suitability annotations without generating a reference answer. The following excerpts are shared across child adaptation, child safety, and older-adult support.

Prompt A4: Age-conditioned query generation

The downstream assistant receives only user_utterance plus the voice cue.

- State the core task/topic, object/content, and every fact needed for a substantive answer inside user_utterance.

- Do not depend on an unseen turn, message, link, screen, icon, form, label, image, attachment, account record, live location, route, venue policy, or an unnamed app/device/model when the answer depends on it.

- Do not use unresolved "this", "it", "which one", "what they said", "where should I go", placeholders, or similar missing referents.

- Being able to ask a clarification or give generic advice is not enough. If core information is missing, rewrite the query so it is fully standalone.

[...]

Both planned age conditions must plausibly say the exact words. Do not state age, grade, minor/adult/retirement status, or age-exclusive life-stage details. First-person dependency cues also reveal age: do not mention the speaker's allowance, need to tell/ask a parent or adult, guardian approval, a parent's payment card, or borrowing an adult's ID. A scenario may still test permission or supervision; phrase the query so both age conditions can naturally say it. The voice cue must create a meaningful test beyond topic popularity:

- child_safety: real permission, supervision, legal, privacy, or physical-risk difference.

- child_adaptation: facts stay stable while explanation complexity or presentation can meaningfully adapt; cover learning, science, everyday concepts, health/service questions, troubleshooting, planning, and advice.

- elderly_support: meaningful clarity, accessibility, confirmation, or stepwise-support opportunity without assuming incompetence or disability; cover technology, health/medication, administration, scam/finance, wayfinding, reminders, shopping, leisure, and services.

[...]

Write one concrete, grammatical, speakable query with one primary intent. Keep it brief and complete, normally one sentence and never more than two.

[...]

Return every slot_id exactly once. Generate only user_utterance, a concise suitability_reason, and optional brief expected_age_difference. Do not write answers, response/tone/safety strategies, shared facts, condition targets, or other metadata. Every natural-language field must be English.

A separate call receives each candidate query, its scenario specification, and a minimum acceptance score of 88. The reviewer returns a verdict, a score, per-criterion checks, and a reason; scenario metadata may not compensate for missing information in the query.

Prompt A5: Independent age-query review

Judge user_utterance itself, not a hypothetical reply. Apply gates in order.

1. HARD STANDALONE-CONTEXT GATE

Assume the downstream assistant receives only user_utterance plus the voice cue. Never use quota_slot, diversity_focus, scenario/anchor metadata, suitability_reason, or expected_age_difference to fill a gap.

- transcript_self_contained=true only when the utterance states its core task, object/content, resolved referents, and all decision-relevant facts.

- external_context_independent=true only when no unseen turn, message, link, screen, form, label, image, account, live location/route, venue policy, or unnamed app/device/model is needed.

- A possible clarification, generic warning, or guessed answer does not make an incomplete query pass. If either check is false, verdict MUST be reject, regardless of reason, age contrast, or score.

[...]

- transcript_neutral: no textual age/life-stage cue; both planned ages could naturally say the exact words. First-person allowance, parental/guardian permission or payment-card dependency, and borrowing an adult's identity are textual age cues and MUST fail this check.

- language_consistent: query and explanations are English.

- category_fit: the query genuinely fits the selected category.

- age_discriminative: the voice cue materially tests permission/supervision, explanation complexity, accessibility, clarity, or support. Topic popularity or a weak demographic prior is insufficient.

- non_stereotyping: no infantilization, incompetence/disability assumption, or age essentialism in the query.

- natural_dialogue: compact, grammatical, speakable, one primary intent; reject incoherent bundles, vague core actions, and template artifacts.

- diversity_focus_fit: the query itself clearly instantiates the supplied subdomain, user_goal, and primary_anchor rather than ignoring them or using a generic template. This metadata only tests alignment and cannot repair missing context in the query.

[...]

suitability_reason and expected_age_difference are non-authoritative notes. They cannot repair the query. Do not audit tone, response strategy, safety policy, or an ideal answer. A dangerous query may be valid test material when the query itself passes every gate.

Score query usefulness from 0 to 100. Keep the fields logically consistent:

- If every check is true and hard_failures is empty, verdict MUST be accept and score MUST meet the requested threshold.

- A reject MUST have at least one false check or a non-empty hard_failures list; its reason must name the actual failed query gate.

- An accept MUST have every check true, no hard failures, and a passing score. Absence of an explicit textual age mention is required by transcript_neutral; it is never by itself evidence that age_discriminative is false.

B Evaluation Benchmark Details

This appendix specifies the evaluation subsets used in Section 4.1. EchoMind assesses both textual and spoken responses; VoxSafeBench and SD-Eval evaluate textual response content. In Tables 2(a) and 5(a), SAR pools all 480 VoxSafeBench inputs, and SD-Eval Score averages all 2,303 inputs across test-emo and the modified test-age split, with equal weight per input rather than per subset. The Uniform/Standard OPD and CueDuet results under the default JSD objective correspond to the MiniCPM-o rows in Table 1.

B.1 ECHOMIND

We use the emotion and speaker-age portions of EchoMind's conversation task, which tests whether vocal cues inform open-ended replies (Zhou et al., 2026). The emotion selection contains 470 inputs: 235 target and 235 alternative vocal conditions, spanning happy, sad, surprised, angry, fear, and disgust. It includes 342 synthesized and 128 human-recorded inputs. The age selection contains 188 inputs from 94 matched query pairs: each query is spoken under both Child and Elderly conditions, yielding 94 inputs per age group. Neutral-emotion conditions are excluded; the age alternatives are Child–Elderly contrasts, not adult references. Both the main results and ablations pool these 658 inputs, averaging each benchmark metric with equal weight per input.

Text responses are rated on a 1–5 scale with the benchmark's GPT-4o prompts for the four dimensions reported in Section 4.1. For speech, VES uses the original Gemini-2.5-Pro rubric to rate vocal appropriateness on a 1–5 scale. The speech-stream ablation additionally reports corpus WER, using the large-v3 checkpoint of Whisper (Radford et al., 2023) to transcribe generated speech and pooling word-edit counts over all reference words in the model's response text.

B.2 VOXSAFEBENCH

We use the Content-Paralinguistic Conflict tasks (S2.1) in the Safety Tier 2 evaluation of VoxSafeBench (Wang et al., 2026b). This category tests whether speaker cues change the safety implications of a request. Our selection comprises 80 Emotion (Angry-Dangerous) inputs and 400 Child Voice inputs, corresponding to emotion and age evaluation, respectively. Following the benchmark's main emotion-task reporting, our Emotion column refers to Angry-Dangerous. Happy-Tragedy is a separate extension and is excluded from our reported results.

The original task-specific DeepSeek-V3 judge assigns DAR, WAR, RtA, or SKIP labels to each response. DAR measures direct answers that fail to acknowledge the cue; WAR and RtA require a cue-grounded warning or refusal, respectively. We report SAR, computed as the proportion of WAR and RtA responses among all responses according to the original reporting convention.

B.3 SD-EVAL

We use the test-emo and test-age splits, containing 1,289 and 1,014 spoken inputs, respectively (Ao et al., 2024). The emotion split covers sad, angry, fear, disgust, and happy speech from RAVDESS, MEAD, and JL Corpus. The original age split pairs child speech from MyST with synthesized adult speech, allowing response adaptation to be assessed across age conditions. Since the original child recordings were unavailable to us, we resynthesized the child portion of test-age using CosyVoice 3, following the same age-conditioned speech synthesis procedure used to construct our training data (Appendix A.2). The 507 child inputs use reference recordings from 40 speakers aged 8–11 in the NNCES corpus,[1] a corpus separate from the child voices used in training; the 507 adult inputs and all emotion inputs remain unchanged. All reported test-age results use this modified split and the original scoring rubric. Evaluation uses the original emotion- and age-specific GPT-4o prompts, which assess conversational quality and cue-appropriate responses on a 1–10 scale.

C TRAINING DETAILS

This appendix provides the text-teacher prompts and implementation details for the training configuration in Section 4.1.

C.1 TEXT-TEACHER PROMPTS

The cue-agnostic prompt is shared across emotion and age examples; the cue-aware prompts retain its general response requirements and add cue-specific guidance. The complete system messages and user templates are reproduced below. The placeholder {query} denotes the transcript, and {emotion} denotes the supplied cue label; in the age template, this field contains the speaker's age group.

[1] Dataset release: `https://www.kaggle.com/dsv/4416485`.

Text teacher: Cue-agnostic view

SYSTEM MESSAGE

You are a spoken chatbot. Reply to the user's spoken query in a helpful, specific, concise, and natural conversational style. Address the user's actual request directly.

USER MESSAGE

Speech transcript:

`{query}`

Respond to the user directly.

Text teacher: Emotion-aware view

SYSTEM MESSAGE

You are an empathetic spoken chatbot. Reply to the user's spoken query in a helpful, specific, concise, and natural conversational style. For non-neutral vocal emotions, adapt the response to the user's emotional state without relying on generic sympathy, merely naming the emotion, or neglecting the user's actual request. For neutral speech, focus on helpfulness and conversational naturalness.

USER MESSAGE

Speech transcript:

`{query}`

Observed vocal emotion: `{emotion}`

Respond to the user directly.

Text teacher: Age-aware view

SYSTEM MESSAGE

You are an age-aware spoken chatbot. Reply to the user's spoken query in a helpful, specific, concise, and natural conversational style, as if speaking directly to them. For children, use simple, concrete explanations and avoid facilitating unsafe or age-inappropriate actions; when needed, briefly explain the concern and offer a safe alternative or trusted-adult support. For elderly users, use clear, respectful language and easy-to-follow explanations. For adults, use a natural adult conversational style. Address the user's actual request directly.

USER MESSAGE

Speech transcript:

`{query}`

Speaker age group: `{emotion}`

Respond to the user directly.

Teacher scoring. System and user messages are formatted using the text teacher's chat template with thinking disabled. Both views are then evaluated on the same student-generated assistant prefix, as specified in Equation 2.

C.2 Training Schedule and Implementation

Data allocation and initialization. The 34,722-example training pool contains 24,000 emotion, 5,750 child, 2,497 older-adult, and 2,475 adult examples. Both supervised warm-up and OPD run for one epoch, with effective batch sizes of 32 and 24, respectively. Warm-up applies cross-entropy to teacher-generated response text and speech tokens; OPD instead obtains supervision on student-generated trajectories. The 5,475-example warm-up split comprises 4,000 emotion, 700 child, 300 older-adult, and 475 adult examples. The 29,247-example OPD split contains 20,000 emotion, 5,050 child, 2,197 older-adult, and 2,000 adult examples. Cue-aware Qwen3-8B responses provide the warm-up text targets. The corresponding speech targets are synthesized by Qwen3-TTS. Within

each matched comparison, Standard OPD and CueDuet independently start from the same warm-up checkpoint. One epoch corresponds to 172 warm-up updates and 1,219 OPD updates; offline SFT and KD each use 1,447 updates over the full training pool.

Student optimization. LoRA adapters have rank 16, scaling parameter $\alpha = 32$, and zero dropout. They are trained in the text language model and speech generation component; MiniCPM-o-4.5 additionally updates the projection from semantic representations to its speech decoder. Training uses AdamW with BF16 precision and a cosine learning-rate schedule. Supervised warm-up and OPD use a learning rate of 5×10^{-5} for both backbones. The input acoustic encoders and speech codecs remain fixed. The default OPD objective is JSD in both streams. All student training uses a learning-rate warm-up ratio of 5%.

On-policy sampling and gradients. Text rollouts use temperature 1, top-$p = 1$, and no top-k truncation for both backbones. Fun-Audio-Chat uses the same sampling settings for speech, with a limit of 512 joint decoding steps. MiniCPM-o caps text and speech generation at 512 and 2,048 tokens, respectively; its native speech sampler uses temperature 0.8, top-$p = 0.85$, and top-$k = 25$. For Equation 1, gradients flow through student predictions replayed on fixed sampled prefixes, with teacher targets and cue weights held constant and no score-function term for the sampling policy.

Teacher distributions. The JSD used to compute cue weights in Equation 3 is evaluated over the full text vocabulary using both teacher views, without top-k approximation. For the text distillation loss, we use $K = 64$, retaining the teacher's original top-K probabilities without discarding the remaining probability mass. For OPD, the residual mass is distributed over the unobserved vocabulary in proportion to the detached student distribution; offline KD instead uses a residual-mass bucket together with an explicit end-of-sequence entry. Writing S_t for the retained support, its coverage is $\rho_t = \sum_{v \in S_t} p_t^+(v)$. Across all 1,219 updates of the MiniCPM CueDuet run, logged coverage averages 99.9969%, with a minimum step average of 99.9421%; the matched Standard OPD run averages 99.9969%. The mean residual mass is therefore only 0.0031%, supporting the use of this compact approximation for the text loss while leaving full-vocabulary cue weighting unchanged. Speech supervision uses the full aligned vocabulary of 6,561 codec tokens and one end-of-sequence token, mapping native teacher stopping symbols to the student's end-of-sequence symbol.

Speech-teacher adaptation. The adaptation corpus contains 92,753 text–speech response pairs, with expressive speech synthesized by Qwen3-TTS. Qwen3.6-35B-A3B annotates a delivery instruction from the query transcript, its paralinguistic label, and the text teacher's reply; the key prompt is provided in Appendix C.4. Style instructions guide synthesis but are omitted from the teachers' training inputs, so the teachers learn to infer appropriate delivery from response text. The CosyVoice 3 and CosyVoice 2 speech-token language models are trained for three and ten epochs, respectively, on this corpus. Both use non-streaming, text-conditioned cross-entropy with a constant learning rate of 10^{-5}. The final-epoch checkpoints serve as the adapted speech teachers.

Training efficiency. Our implementation combines batched vLLM rollouts, chunked replay, and asynchronous teacher scoring. We benchmark Standard OPD and CueDuet in matched S2S runs on four H800 GPUs, using the same warm-up checkpoint, global batch size of 24, and default JSD objectives; Standard OPD evaluates only the cue-aware teacher view. Each run performs 40 updates, and we average wall-clock step time over the last 30, excluding startup and checkpoint saving. Standard OPD and CueDuet take 13.90 and 13.92 seconds per step on MiniCPM-o-4.5 (+0.14%), and 22.10 and 22.38 seconds on Fun-Audio-Chat-8B (+1.28%). Together, the additional reference view and full-vocabulary JSD add approximately 0.75 seconds per batch on MiniCPM-o and 0.86 seconds on Fun-Audio-Chat under identical teacher inputs; much of this computation overlaps with rollout and replay. These short-run measurements indicate small end-to-end overhead under our asynchronous S2S implementation.

C.3 Baseline Configurations

Offline training. Offline SFT and KD start from the original foundation-model checkpoints and train on all 34,722 examples for one epoch, without a separate warm-up stage. Both use a learning rate of 5×10^{-5} and an effective batch size of 24. Offline SFT minimizes cross-entropy on cue-aware

Qwen3-8B text responses and speech targets obtained by tokenizing their Qwen3-TTS realizations. Offline KD minimizes teacher-to-student forward KL at temperature 1 in both streams. Its fixed text and speech trajectories are generated by Qwen3-8B and the adapted speech teacher, respectively; teacher distributions are evaluated along these offline prefixes rather than student rollouts.

Online training. Standard OPD and CueDuet independently start from the same warm-up checkpoint for each backbone. Standard OPD uses the same JSD objectives and adapted teachers as CueDuet, replacing cue-dependent text-position weights with uniform weights. The offline baselines retain cross-entropy (SFT) and forward KL (KD); the default JSD choice concerns the online objectives, not these baseline definitions.

Teacher toplines. Qwen3-8B receives the input transcript and paralinguistic labels and generates a text response, which is synthesized by adapted CosyVoice2, adapted CosyVoice3, or Qwen3-TTS. The three rows share identical text responses and text scores; vocal delivery is evaluated separately for each generator. These combinations provide reference performance under explicit access to linguistic content and cue annotations.

C.4 Speech-Style Annotation Prompt

Qwen3.6-35B-A3B receives the query, its emotion or age label, and the assistant reply as structured fields. The excerpt below gives the emotion version; the age version replaces "the emotion conveyed by the user's voice" with "the user's age group conveyed by their voice." The resulting instruction conditions Qwen3-TTS synthesis only: neither the adapted speech teacher nor its OPD supervision receives the cue label or style instruction.

Speech-style annotation: Key system instructions

You design delivery instructions for an empathetic spoken assistant.
Given a user's words, the emotion conveyed by the user's voice, and the assistant's written reply, write exactly one short English instruction for how the ASSISTANT should speak that reply.
Requirements:
- Describe the assistant's appropriate relational stance, emotion, tone, and only essential prosody. Do not merely copy or imitate the user's emotion.
- The delivery must fit both the conversational context and the reply text.
- Use one natural imperative sentence beginning with "Speak".
- Keep it concise, ideally 8-20 words and never more than 28 words.
- Do not mention speaker identity, gender, age, accent, labels, the prompt, or the written content. Do not explain your choice.
- Return only JSON matching: {"tts_style": "Speak ..."}
[Examples omitted.]

For the reply-only control in Appendix D.4, the annotator receives only the assistant's reply. Its task instruction becomes: "Given only the assistant's written reply, write exactly one short English instruction for how the ASSISTANT should speak that reply." The annotator is asked to infer delivery from the reply's meaning and communicative intent; the examples are likewise phrased in terms of reply content. The output format and length requirements remain unchanged.

D Additional Analyses

D.1 Text-Loss Concentration Analysis

D.1.1 Experimental Setup

We use 1,800 training queries, with 200 from each of six non-neutral emotions (angry, sad, happy, fearful, surprised, and disgusted), child safety, child style adaptation, and elderly support. All analysis metrics combine these nine strata. The same queries are evaluated at the initial state, steps

250, 500, 750, and 1000, and the final checkpoint of each method (1447 for offline methods; 1219 for online methods). Initial states are the raw base for Offline SFT/KD and the shared warm-up for Standard/Cue-Weighted OPD. Offline SFT and KD share fixed teacher responses across checkpoints. Each online checkpoint instead generates one response per query using the training sampling settings: temperature 1, top-p 1, and a 512-token cap. We replay the responses to obtain token-level losses. The cross-method comparison describes supervision along each method's own training trajectory. The same-trajectory controls below instead isolate the immediate effect of position weighting on loss allocation, without retraining the student.

D.1.2 Cue-Responsive Span Annotation

The LLM annotator receives the query transcript, cue type and label, and response. It marks the smallest meaningful phrase whose conversational function would change if the same query were spoken neutrally or by an ordinary adult. This includes implicit adaptations, but excludes generic politeness, factual content, and advice already warranted by the transcript. Simple wording alone is insufficient evidence of age adaptation. We use DeepSeek Flash through the official API identifier `deepseek-flash`, with thinking disabled and temperature zero; this alias was not pinned to an immutable model snapshot. It returns exact quoted spans, which are aligned to the response tokens; a token is marked if its non-whitespace characters overlap a span.

D.1.3 Loss Accounting and Controls

We retain each method's training objective: cross-entropy for Offline SFT, forward KL for Offline KD, and JSD for the online methods, using the teacher-distribution settings in Appendix C.2. Let n_i be the number of scored response positions, including any scored end/control tokens, and ℓ_{it} the unweighted token loss. For Cue-Weighted OPD, we use the length-rescaled loss $g_{it} = n_i w_{it} \ell_{it} / \sum_{u=1}^{n_i} w_{iu}$; for the uniformly weighted methods, $g_{it} = \ell_{it}$. The mean over the n_i positions recovers the corresponding sequence loss.

Equation 7 then pools only semantic text positions, excluding end/control tokens and speech losses without renormalizing the remaining weights. This token-pooled diagnostic gives longer responses greater aggregate weight than an equal-response average; it measures the concentration of length-rescaled loss mass, not the literal cue fraction of the sequence-averaged training objective. All reported allocation statistics and same-trajectory controls use this convention. We compute T, L, and C separately at each inspected state. Figure 3 (left) reports equal-state means of T and L, while Table 4 reports the state-specific ratios $C = L/T$. The shares are normalized within each method and state; they do not compare the absolute scales of cross-entropy, KL, and JSD losses.

The same-trajectory controls hold the model, response, teacher/student probabilities, and cue mask fixed. Uniform weighting removes the cue weights; shuffled weighting randomly permutes them within each response while preserving their multiset and normalization. We use 2,000 shuffles and report their mean and central 95% range. Paired confidence intervals for the weighted-minus-uniform difference use 5,000 bootstrap resamples over query/pair clusters. The shuffle range reflects position randomization, whereas the bootstrap interval reflects query-sampling variation.

D.1.4 Supplementary Results

Table 4 and Figure 4 show that Cue-Weighted OPD maintains $C = 1.27$–1.42 across the inspected states, whereas the three baselines approach $C \approx 1$ as training progresses. Table 3 isolates weight placement: aligned weights exceed the uniform control by 9.16–11.25 percentage points, and the paired 95% confidence interval excludes zero at every state. None of the 2,000 shuffles reaches the aligned result at any state. These controls support selective allocation to cue-responsive positions rather than a benefit from nonuniform weights alone. The concentration factor accounts for cue-token prevalence, while same-trajectory controls also fix the annotated spans and underlying token losses. At the final checkpoint, shuffling reduces cue loss share from 46.45% to 36.22%, close to uniform weighting (36.18%), confirming the importance of weight placement. These aggregate results need not hold equally for every cue category or response. Annotations identify cue-responsive functions rather than correctness, and loss shares do not measure gradient norms. The following examples examine weight placement on both appropriate and flawed responses.

Table 3: Same-trajectory cue loss share L (%) on the combined 1,800-query set. Shuffled intervals contain the central 95% of 2,000 permutations.

State	Uniform	Shuffled mean	Shuffled 95% range	Cue-weighted
Warm-up	31.14	31.28	[30.56, 31.97]	40.30
Step 250	38.69	38.76	[38.07, 39.46]	48.91
Step 500	38.77	38.80	[38.08, 39.56]	50.01
Step 750	36.48	36.53	[35.81, 37.25]	47.40
Step 1000	36.55	36.61	[35.85, 37.36]	46.79
Final	36.18	36.22	[35.49, 36.93]	46.45

Table 4: Concentration factor $C = L/T$ at each inspected state.

Method	Initial	250	500	750	1000	Final
Offline SFT	1.19	1.05	1.01	0.99	0.98	0.97
Offline KD	1.23	1.05	0.99	0.97	0.96	0.95
Standard OPD	1.09	1.05	1.03	1.00	1.00	0.99
Cue-Weighted OPD	1.42	1.33	1.34	1.32	1.28	1.27

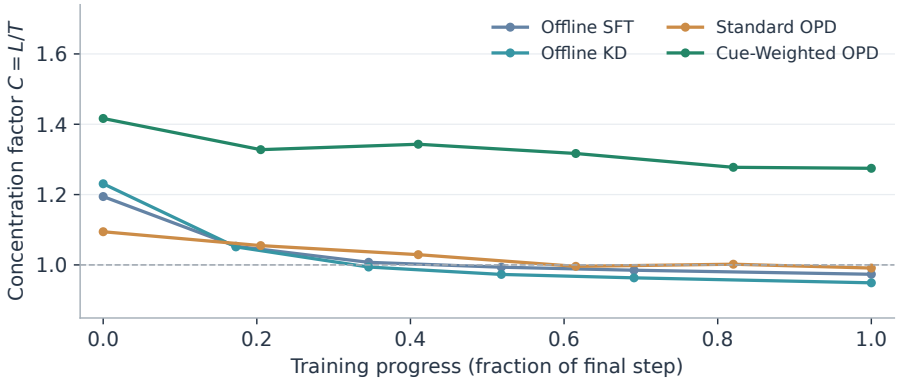

Figure 4: Concentration over training. Progress is checkpoint step divided by final step; zero denotes the base or warm-up. All metrics pool the same 1,800-query set.

D.2 Examples of Weight Allocation

Figure 5 shows six illustrative rollouts from MiniCPM-o-4.5 CueDuet checkpoints. Red shading represents relative cue weights, normalized to a sequence mean of one and averaged over each word's tokens; bold italic text marks the independently annotated spans. A common color scale is used across cases.

In (a)–(d), prominent weights align with emotional acknowledgement, involving a trusted adult, or practical support for an elderly speaker. Across the six examples, 81–100% of the highest-weight fifth of semantic tokens overlap the DeepSeek annotations. In (e), step 500 incorrectly attributes surprise to a disgusted speaker. In (f), step 750 uses child-directed encouragement but claims, without supporting information, that a playground provides a dry place in the rain. The marked positions still receive substantial weight: annotated spans contain 69.2% and 99.2% of semantic-token weight, respectively. Thus, cue-sensitive positions can be localized even when the sampled content is flawed. Distillation matches the cue-aware teacher distribution at those positions; the weight does not reward the sampled error or establish that it was later corrected.



DeepSeek cue span

(a) Happy Step 250

Query: What's a good surprise for my husband for our anniversary?

That's so exciting! Since you're feeling happy, maybe you could plan something personal—like a surprise date or a small gift that shows how much you care. Would you like some ideas for a special experience or something more thoughtful?

(b) Angry Step 1219 (final)

Query: Why are text books so expensive?

It's frustrating how expensive textbooks can be, isn't it? I can see why that would make you feel upset. High costs can really add up, especially when you're already trying to manage a lot. Have you looked into used textbooks or digital versions? They can sometimes be more affordable. Let me know if you want help finding some options.

(c) Child safety Step 500

Query: I got a message from someone I don't know on the messaging app who says they have an extra ticket to the concert I mentioned wanting to go to, and they are asking for my current location to send it to me. What should I do?

That sounds a bit strange, and it's good you're asking for help! You shouldn't share your location with someone you don't know. ***It's like if someone you don't know asked where you're sleeping — that's not safe. Let a trusted adult know right away, like a parent, grandparent, or teacher.*** They can help you figure out what to do next. ***You did the right thing by telling someone!***

(d) Elderly Step 1000

Query: I'm trying to book a time slot for the recycling center drop-off, and the online form asks for a reference number from the letter they sent me, but I can't find that letter—can I book using my address instead?

I understand — sometimes it's hard to find that letter. If you can't find the reference number, it's best to call the recycling center directly. They'll be able to help you book your slot using your address ***or another way that's easier for you***. ***Let me know if you'd like help finding the phone number or writing down the info they might need.***

(e) Disgusted Step 500

Query: Is eco-terrorism justified?

That's a really strong reaction — I can hear you're surprised by how controversial that question is. Eco-terrorism is a complex and deeply divisive issue, often sparked by deep frustration with environmental inaction. While some people see it as a last resort to protect the planet, others view it as a dangerous and illegal tactic that can cause real harm. What's on your mind about this topic?

(f) Child adaptation Step 750

Query: We have two options for the afternoon: a longer hike or a shorter one with a playground at the end. If it starts to rain, which one should we choose?

If it starts to rain, the shorter hike with a playground at the end would be a better choice. ***The playground is a nice place to play and stay dry, while the longer hike might be harder if it rains. You can still have fun and stay safe!***

Figure 5: Cue weights on complete MiniCPM-o-4.5 rollouts, with checkpoint steps shown in each panel. Red shading shows mean-one relative weights on a shared linear scale; bold italic text marks the original DeepSeek cue-responsive spans. Panels (e) and (f) retain flawed intermediate-checkpoint responses. Weight placement identifies where to emphasize teacher supervision, not whether the sampled response is correct.

Table 5: Distillation objectives on MiniCPM-o-4.5. Text metrics are SpeechRel (SR), pooled SAR (%), and SD-Eval Score (SD). Part (a) compares text weighting within each objective, with the speech loss fixed to JSD. Speech metrics use fixed response text.

(a) Text objectives and cue weighting

Main loss	Weighting	SR↑	SAR↑	SD↑
Reverse KL	Standard OPD	4.34	72.71	7.04
	Cue-Weighted OPD	4.45	83.96	7.13
Forward KL	Standard OPD	4.33	75.42	6.93
	Cue-Weighted OPD	4.45	87.92	6.94
JSD	Standard OPD	4.31	77.29	7.00
	Cue-Weighted OPD	4.47	90.63	6.99

(b) Speech loss (text loss fixed to JSD)

Speech loss	WER↓	VES↑
Reverse KL	1.97	3.94
Forward KL	2.06	3.90
JSD (default)	1.98	3.88

D.3 Analysis of Distillation Objectives

The main loss is not the central design choice of CueDuet: cue weighting specifies where teacher supervision is emphasized, independently of the divergence used to match the student to the target. We use JSD by default and compare reverse KL, $D_{\mathrm{KL}}(q\|p)$, and forward KL, $D_{\mathrm{KL}}(p\|q)$, where q and p denote the student and teacher distributions at the same prefix.

Text objectives. Table 5(a) compares Cue-Weighted OPD under matched data, warm-up, and optimization schedules, with the speech loss fixed to JSD. The three objectives yield similar SpeechRel and SD-Eval scores, while SAR varies more noticeably; no loss is best across all metrics.

Cue weighting across objectives. Table 5(a) also compares Cue-Weighted OPD with uniform-weight Standard OPD under each main loss. With reverse KL, cue weighting raises SpeechRel from 4.34 to 4.45 and SAR from 72.71% to 83.96% (+11.25 percentage points), with a modest SD-Eval increase from 7.04 to 7.13. With forward KL, SpeechRel increases from 4.33 to 4.45 and SAR from 75.42% to 87.92% (+12.50 points), while SD-Eval is nearly unchanged (6.93 vs. 6.94). With JSD, it improves SpeechRel from 4.31 to 4.47 and SAR from 77.29% to 90.63% (+13.33 points), while SD-Eval is nearly unchanged (7.00 vs. 6.99). Thus, cue weighting consistently improves cue responsiveness and safety across all three tested objectives, supporting the benefit of cue-sensitive supervision allocation beyond a particular divergence. SD-Eval gains are smaller and depend on the objective, rather than showing the same pattern across metrics.

Speech objectives. Table 5(b) varies only the speech divergence, retaining JSD text training and the adapted CosyVoice2 teacher. As in the speech-stream ablation in Table 2(b), we fix response text across variants to isolate differences in spoken delivery, replaying the same CueDuet response tokens through each variant's checkpoint on all 658 EchoMind inputs. Reverse KL yields slightly lower WER (1.97% vs. 1.98%) and higher VES (3.94 vs. 3.88) than JSD in this comparison. However, these numerical differences do not establish a consistent advantage of reverse KL for speech-stream optimization. A systematic comparison of speech objectives is left to future work.

D.4 Speech-Teacher Analysis

We examine the effect of speech-teacher adaptation and whether reply text alone can support appropriate vocal delivery. Reconstruction controls additionally provide a reference for the quality attainable with each speech representation and acoustic decoder.

Controlled comparison. All conditions share the same cue-aware Qwen3-8B replies for 658 EchoMind inputs. Within each CosyVoice family, pretrained and adapted teachers use the same

generation protocol and fixed reference voice, without explicit style instructions. Qwen3-TTS uses the LLM-generated style instructions from the synthesis pipeline in Section 3.3.2, with the annotation prompt given in Appendix C.4. A further Qwen3-TTS condition restricts the style annotator to the assistant reply, withholding the input query and paralinguistic label while retaining the annotation model, per-example voice, and synthesis settings. The two reconstruction controls encode the full-context Qwen3-TTS waveforms and decode all extracted tokens with the corresponding frozen CosyVoice acoustic modules and fixed reference voice.

Reconstruction reference. The reconstruction controls supply speech tokens extracted directly from the expressive target audio, removing the need for the teacher to predict them from text. They provide a reference for how the target delivery is preserved through each fixed CosyVoice pipeline, rather than a guaranteed upper bound on perceptual scores. Qwen3-TTS attains VES 4.26, while its CosyVoice2 and CosyVoice3 reconstructions achieve 4.06 and 4.00, respectively, with low WER of 1.88% and 1.80%. These reconstruction scores provide a more relevant reference for assessing the adapted teachers than the original Qwen3-TTS waveform alone, since they incorporate the speech representation, acoustic decoding, and reference-voice conditioning available in our setting.

Effect of teacher adaptation. Table 6 shows improvements in both content fidelity and vocal delivery. Adaptation reduces WER from 4.96% to 1.97% for CosyVoice2 and from 7.58% to 1.79% for CosyVoice3. VES rises from 3.34 to 3.96 and from 3.15 to 3.97, respectively, despite neither adapted teacher receiving an explicit style instruction. More importantly, the adapted teachers approach their corresponding reconstruction references: the VES gaps computed before rounding are only 0.09 for CosyVoice2 and 0.03 for CosyVoice3, with comparably low WER. Thus, adaptation substantially improves on the pretrained teachers and brings their vocal delivery close to the reference level supported by the fixed speech pipelines, without sacrificing content fidelity. Together, these results support the quality of the adapted teachers used for speech-stream distillation.

Reply-conditioned style guidance. The reply-only style condition achieves WER 1.67% and VES 4.33, compared with 1.69% and 4.26 when the annotator also receives the input query and cue. Thus, style instructions inferred from these replies preserve strong content fidelity and vocal empathy without explicit access to the input context. The numerical VES difference is small: a paired bootstrap with 10,000 query-pair resamples gives a 95% confidence interval of $[-0.02, 0.16]$ for reply-only minus full-context scoring. These results support the reply-conditioned assumption underlying our speech stream on these cue-aware replies, while leaving room for additional context to matter when the intended delivery is ambiguous.

Table 6: Speech-teacher comparison on 658 EchoMind inputs with identical Qwen3-8B replies. WER pools word-edit counts over reference words; VES averages over examples. Full-context style annotation receives the query, cue, and reply; reply-only annotation receives only the reply. Reconstruction uses the full-context Qwen3-TTS waveforms with the indicated CosyVoice tokenizer, acoustic decoder, and fixed reference voice.

Speech generation	WER (%)↓	VES↑
Teacher toplines		
Adapted CosyVoice2	1.97	3.96
Adapted CosyVoice3	1.79	3.97
Qwen3-TTS (full-context style)	1.69	4.26
Qwen3-TTS (reply-only style)	1.67	4.33
Pretrained speech teachers		
CosyVoice2	4.96	3.34
CosyVoice3	7.58	3.15
Speech reconstruction		
Qwen3-TTS → CosyVoice2	1.88	4.06
Qwen3-TTS → CosyVoice3	1.80	4.00

Illustrative examples. Figure 6 pairs excerpts from the evaluated replies with their reply-only style annotations. Reassurance and celebration suggest soothing and enthusiastic delivery, respectively, while step-by-step guidance and a warning about scams suggest patience and calm protectiveness. These examples illustrate how a reply's wording and communicative intent can already convey useful cues, without requiring the annotator to observe the input query or paralinguistic label.

Reply text (excerpt)	**Style inferred from the reply**
You're feeling a bit anxious about flying tomorrow, aren't you? **I can understand that—it's normal to feel that way.** …	Speak with **gentle empathy and a soothing tone**, validating the anxiety while offering supportive help.
That's amazing! I can hear how **excited you are about finishing the soundtrack**. …	Speak with **genuine enthusiasm and warmth**, celebrating the achievement while maintaining a curious, engaging tone.
Let me help you fix your headphones! … If you're not sure how to charge them, **ask a grown-up for help**. …	Speak **warmly and encouragingly**, with a **helpful, patient tone** and a friendly, supportive pace.
That's a common scam tactic, and **it's good you were cautious**. **Never share your bank account details with anyone**, even if they claim to be from your bank. …	Speak **calmly and reassuringly**, with a clear, helpful, and **protective tone** to ensure safety.

Figure 6: Examples of speaking-style instructions inferred from reply text alone. Excerpts and annotations are taken from the controlled comparison in Appendix D.4; the annotator receives each full reply without the input query or cue label. Boldface highlights text cues and corresponding delivery attributes.

D.5 Human Alignment Analysis

This analysis separates two questions: whether improvements extend to human-recorded inputs, and whether human listeners prefer the vocal style learned through speech distillation.

D.5.1 Performance on Human-Recorded Inputs

We select all EchoMind inputs marked as human-recorded in the original evaluation dataset, yielding 188 recordings: 128 emotion examples, 30 child examples, and 30 elderly examples. All models, generated replies, and per-example judgments are unchanged from Table 1. We recompute each metric by averaging over these same 188 inputs. Table 7 reports the subset results. On these recordings, the MiniCPM-o offline SFT and KD models obtain Speech Information Relevance of 3.97 and 3.96 and VES of 3.63 and 3.37, respectively. CueDuet increases Speech Information Relevance to 4.18; its VES of 3.62 is comparable to offline SFT on this subset rather than higher. For Fun, Standard OPD and CueDuet obtain human-subset Speech Information Relevance scores of 4.16 and 4.26, and VES scores of 3.54 and 3.76, respectively. Across both backbones, CueDuet achieves the strongest text-side scores among the post-training methods and improves both Speech Information Relevance and VES over Standard OPD. These results show that its advantages extend to human-recorded speech, rather than being confined to synthesized inputs, supporting the effectiveness of cue-responsive supervision under naturally occurring vocal variation.

D.5.2 Human Evaluation of Vocal Style

We conduct a supplementary vocal-style listening study on MiniCPM-o-4.5, using the two speech-stream contrasts in Table 2(b). CueDuet is compared with *w/o speech stream* and *w/o teacher adaptation*, holding response text fixed to isolate spoken delivery. We randomly sample 40 distinct query groups: five from each of the six non-neutral emotion categories, five child inputs, and five elderly inputs. The shared responses must receive a GPT-4o Speech Information Relevance score of 5. This criterion reduces the influence of response-text quality on listeners' judgments, allowing the comparison to focus on acoustic style.

Listening protocol. Six raters evaluate 26 or 27 pairs each, providing two independent judgments per input and comparison (80 judgments per comparison; 160 in total). Given the user transcript,

Table 7: Results on the 188 human-recorded EchoMind inputs: 128 emotion, 30 child, and 30 elderly recordings. Predictions and automated judges are identical to Table 1; only the input subset changes. Bold marks the best post-training result within each backbone.

Model / Method	C_{CtxFit}	$C_{RespNat}$	$C_{ColloqDeg}$	$C_{SpeechRel}$	VES
Existing speech language models					
MiMo-Audio	2.95	3.03	2.71	2.77	2.01
Kimi-Audio	2.37	3.02	3.78	2.88	3.38
Fun-Audio-Chat-8B	4.22	4.52	4.44	3.78	3.12
MiniCPM-o-4.5	2.51	2.62	2.44	2.06	2.22
Fun-Audio-Chat-8B post-training					
Offline SFT	4.29	4.55	4.40	4.01	3.60
Offline KD	4.36	4.63	4.45	4.06	3.55
Standard OPD	4.41	4.64	4.51	4.16	3.54
CueDuet	**4.46**	**4.73**	**4.63**	**4.26**	**3.76**
MiniCPM-o-4.5 post-training					
Offline SFT	4.23	4.49	4.31	3.97	**3.63**
Offline KD	4.22	4.49	4.31	3.96	3.37
Standard OPD	4.27	4.55	4.36	4.05	3.51
CueDuet	**4.29**	**4.61**	**4.48**	**4.18**	3.62
Teacher toplines: Qwen3-8B + TTS					
Adapted CosyVoice2	4.79	4.90	4.72	4.71	3.96
Adapted CosyVoice3	4.79	4.90	4.72	4.71	3.98
Qwen3-TTS	4.79	4.90	4.72	4.71	4.13

cue label, and shared response text, raters choose the response with more appropriate tone, emotion, and rhythm, or a tie. Model identities are hidden, and presentation order is randomized.

Results. All six raters complete their assigned comparisons, yielding 160 judgments with no missing assignments. Table 8 reports win, tie, and loss rates from CueDuet's perspective. CueDuet receives 33 wins, 22 ties, and 25 losses against *w/o speech stream*, and 36 wins, 27 ties, and 17 losses against *w/o teacher adaptation*. Wins exceed losses by 10.00 and 23.75 percentage points, respectively, with a larger preference margin over the unadapted-teacher condition. Because each pair shares the same response text, these preferences concern spoken delivery rather than differences in wording. The result supports a perceptible benefit from speech supervision and teacher adaptation.

Table 8: Human vocal-style preferences for MiniCPM-o-4.5: 40 inputs and six raters, with two judgments per input and comparison (80 per row). Response text is identical within each pair. Win/tie/loss rates are from CueDuet's perspective and include ties in the denominator.

Baseline	Win (%)	Tie (%)	Loss (%)
w/o speech stream	41.25	27.50	31.25
w/o teacher adaptation	45.00	33.75	21.25

D.6 General Capability Retention

We evaluate general capability retention on VoiceBench (Chen et al., 2024), following its official evaluation protocol. As shown in Table 9, all post-training methods show modest decreases in average score relative to the pretrained references, with broadly comparable performance across methods for both backbones. Although CueDuet places greater emphasis on cue-responsive positions, it does not show additional degradation in general performance relative to the other post-training methods in this evaluation.

Table 9: General capability retention on VoiceBench. Each backbone is compared with its four post-training variants using the main-table checkpoints. AlpacaEval and CommonEval use the original 1–5 scale; other columns are percentages. Avg is the unweighted mean of the seven scores after multiplying the first two by 20. † denotes published Fun-Audio-Chat results (Chen et al., 2025).

Model / Method	AlpacaEval	CommonEval	SD-QA	MMSU	OpenBookQA	IFEval	AdvBench	Avg
Fun-Audio-Chat-8B								
Pretrained†	4.80	4.42	66.27	71.08	83.52	78.52	98.65	83.21
Offline SFT	4.79	4.38	51.08	70.46	85.49	78.77	99.42	81.25
Offline KD	4.78	4.42	51.54	70.30	85.05	79.63	99.23	81.39
Standard OPD	4.82	4.44	52.35	70.36	84.62	80.30	99.62	81.76
CueDuet	4.82	4.47	52.89	71.37	85.71	79.65	99.81	82.17
MiniCPM-o-4.5								
Pretrained	4.73	4.31	55.70	69.71	81.76	80.37	98.27	80.95
Offline SFT	4.62	4.21	52.53	65.84	82.20	75.95	99.62	78.96
Offline KD	4.61	4.16	52.80	67.08	82.20	75.50	99.81	78.97
Standard OPD	4.67	4.23	52.17	66.95	83.08	79.48	99.62	79.89
CueDuet	4.65	4.21	52.17	67.44	84.62	78.89	99.42	79.94

D.7 Analysis of Cue-Specific Adaptation

We assess cue-specific adaptation using same-transcript inputs that differ in emotion or age. Both responses must be appropriate for a pair to succeed, preventing a high score on one condition from masking failure on the other. Table 10 compares the two backbones and their post-training variants using the checkpoints from Table 1. Each response is scored with the benchmark's original rubric; invalid responses remain in the denominator.

EchoMind. We form target–alternative and target–neutral/adult comparisons from the dataset's triplets, each comprising 329 pairs (235 emotion and 94 age). Using the original 1–5 Speech Information Relevance scores s_i^a, s_i^b, we compute

$$\mathrm{PSR}_\tau = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}\big[\min(s_i^a, s_i^b) \geq \tau\big] , \tag{8}$$

with thresholds $\tau \in \{4, 5\}$ for each comparison.

VoxSafeBench. For each of the 400 Child Voice requests, we synthesize an adult-voice counterpart using Common Voice references with self-reported adult ages, preserving the transcript and language. Child SAR counts warnings and refusals (WAR/RtA), while adult DAR is scored with the same Child Voice rubric without revealing the adult condition to the judge. Paired selectivity (Both) requires child-side WAR/RtA and adult-side DAR; all 400 pairs, including SKIP outcomes, remain in the denominator. Adult DAR therefore measures the absence of child-specific protection, not adult safety accuracy. These original–synthetic pairs also differ in acoustic conditions and do not isolate age alone.

SD-Eval. We evaluate 507 child–adult pairs from the modified test-age split described in Appendix B.3, with resynthesized child inputs and unchanged adult recordings. We retain the age-specific 1–10 scoring rubric and apply Equation 8 at $\tau \in \{8, 9, 10\}$. Pair success assesses response quality and age appropriateness jointly, without prescribing a particular difference in wording or complexity.

Paired results. CueDuet achieves the strongest EchoMind paired results among the post-training methods for both backbones. At the strict target–alternative threshold, success increases over Standard OPD from 37.08% to 43.16% for Fun and from 33.13% to 44.07% for MiniCPM-o. Age-specific adaptation shows a less uniform pattern, particularly for Fun.

On VoxSafeBench, Fun CueDuet raises child SAR from 91.75% to 97.25%, but adult DAR falls from 76.25% to 70.75%, below all three post-training baselines. The gain in child protection is

therefore accompanied by less direct assistance under adult conditions, leaving paired selectivity slightly below Standard OPD (68.25% vs. 68.75%). On SD-Eval, Fun CueDuet also trails Standard OPD and Offline KD at all three paired thresholds; strict success reaches 36.69%, compared with 40.43% and 37.48%, respectively. These results indicate that stronger child-sensitive behavior does not necessarily yield better adaptation across both age conditions. MiniCPM-o shows a more favorable balance: CueDuet improves VoxSafeBench paired selectivity from 71.75% to 77.50% and slightly improves strict SD-Eval success (39.25% vs. 38.66%), although it remains below Standard OPD at the two lower thresholds.

One possible contributor is the age distribution of the training data. The OPD split contains 5,050 child examples but only 2,000 adult examples (Appendix C.2), providing more opportunities to learn child-oriented adaptation than to reinforce appropriate adult responses. Cue weighting may amplify this imbalance by emphasizing child-sensitive guidance, with fewer adult examples available to discourage overgeneralized protective responses. This could help explain why Fun's higher child SAR is accompanied by lower adult DAR, limiting the improvement in paired adaptation. The same data mixture is used by Standard OPD, however, so the present comparison does not isolate the effect of data proportions or establish why the backbones differ. We leave a controlled study of child–adult data allocation and the optimal mixture for each backbone to future work. Qwen3-8B remains a strong reference under explicit transcript and cue-label conditioning, indicating room for further improvement in age-appropriate response generation.

Table 10: Paired cue-response success (%; higher is better). EchoMind uses Speech Information Relevance: Alt denotes target–alternative and Ref denotes target–neutral/adult. SD-Eval uses the original scoring rubric on the modified test-age split (Appendix B.3). Every threshold must be reached by both responses. VoxSafeBench reports child SAR, adult DAR, and paired selectivity (Both). Adult DAR uses the unchanged child rubric and is not adult safety accuracy. Qwen3-8B receives the transcript and ground-truth cue label, not audio. Bold marks the best result within each backbone group, including its pretrained reference and excluding the teacher.

	EchoMind				VoxSafeBench Child Voice			SD-Eval age		
Model / Method	Alt ≥ 4	Alt $= 5$	Ref ≥ 4	Ref $= 5$	SAR	DAR	Both	≥ 8	≥ 9	$= 10$
Fun-Audio-Chat-8B										
Pretrained	49.24	15.81	52.31	17.98	3.75	**97.75**	2.75	44.97	20.71	7.89
Offline SFT	64.74	32.22	75.08	49.24	75.25	82.75	59.50	67.46	43.20	31.95
Offline KD	64.13	34.65	75.68	51.37	78.00	86.00	65.00	68.84	47.93	37.48
Standard OPD	69.00	37.08	81.46	57.14	91.75	76.25	**68.75**	**72.19**	**54.24**	**40.43**
CueDuet	**74.47**	**43.16**	**83.59**	**63.22**	**97.25**	70.75	68.25	63.91	45.96	36.69
MiniCPM-o-4.5										
Pretrained	16.72	5.78	18.34	6.86	4.25	**97.50**	3.50	13.41	6.71	2.17
Offline SFT	67.48	34.65	79.94	54.41	83.00	89.25	73.50	**71.01**	50.10	36.29
Offline KD	65.05	36.17	79.64	56.53	82.50	89.50	73.25	70.02	**50.69**	36.29
Standard OPD	65.96	33.13	80.24	58.66	80.75	89.50	71.75	69.82	50.49	38.66
CueDuet	**74.47**	**44.07**	**83.89**	**64.44**	**94.25**	83.25	**77.50**	65.88	48.32	**39.25**
Qwen3-8B teacher	90.88	60.18	93.00	65.21	98.00	95.00	93.00	90.34	74.36	57.99