VOXTACT: PRESERVING VOCAL NUANCE FOR ACOUSTICALLY APPROPRIATE RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

The same words can call for different responses when spoken with different affect, yet responding distinctively can also introduce unsupported assumptions. We present VoxTact, a spoken-dialogue framework that learns when vocal nuance should change a reply while preserving factual and role consistency. ToneTrace retains temporally ordered acoustic summaries alongside the existing speech representation. Appropriateness-Relative Policy Optimization (ARPO) evaluates each candidate reply across paired contexts and rewards supported specificity under quality and relation constraints. ToneFoil provides 100,000 training recordings with reviewed response relations; ToneFoil-Bench evaluates strictly matched wording and distinguishes justified adaptation from appropriate shared responses. ToneTrace preserves the cues; ARPO learns their relevance to response suitability; ToneFoil tests the distinction. In matched response-learning comparisons, VoxTact achieves 63.34% bilateral response success, compared with 49.94% for sample-level policy optimization, while attaining 94.10% quality acceptance and 2.90% unsupported assertions. Representation and objective controls distinguish the contribution of acoustic input from additional capacity and expose the cost of rewarding specificity without quality constraints. The results support selective acoustic conditioning rather than indiscriminate emotional differentiation.

1 INTRODUCTION

"The meeting is tomorrow" can express relief or apprehension. An appropriate assistant might offer encouragement in one case and practical support in the other, while preserving the same meeting date. The words constrain what the assistant can claim; the delivery can influence which otherwise valid response is useful. This distinction is central to spoken interaction because accurate transcription and fluent generation do not, by themselves, determine an appropriate conversational emphasis (Figure 1).

Text-based empathetic dialogue relates responses to emotional situations and support strategies (Rashkin et al., 2019; Liu et al., 2021). Speech language models extend this setting by conditioning large language models (LLMs) on audio and producing spoken replies (Zhang et al., 2023; Chu et al., 2024; Zeng et al., 2024). Recent affect-aware models connect vocal perception more directly to response generation. Their progress sharpens a remaining question: how can learning distinguish useful acoustic adaptation from an unnecessarily different or factually unjustified reply?

Two individually acceptable replies need not be interchangeable across vocal contexts. Conversely, a response that fits several contexts can be entirely appropriate. An absolute quality score therefore cannot by itself identify the relevance of the acoustic distinction. Nor does a difference between emotion labels establish that two replies should be preferred in opposite contexts. The learning signal must connect audible evidence to candidate suitability while permitting shared replies and explicit uncertainty.

VoxTact addresses this problem through three connected components. *ToneTrace* preserves temporally ordered acoustic information as additional input vectors. *Appropriateness-Relative Policy Optimization* (ARPO) evaluates the same candidate under paired contexts, rewarding a positive suitability difference only when quality and the reviewed relation support it. *ToneFoil* supplies these relations, and *ToneFoil-Bench* tests their behavioral consequences under strictly matched words. De-

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.

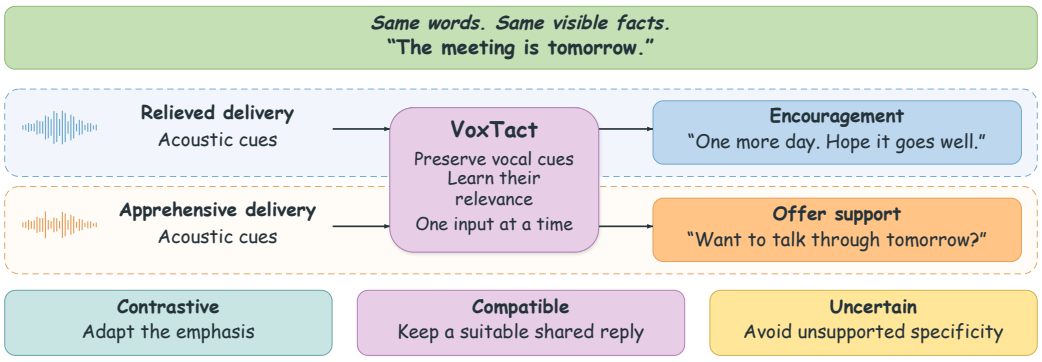

Figure 1: Fixed words and visible facts constrain both replies; vocal evidence can justify different emphasis. Contrastive, compatible, and uncertain relations distinguish warranted adaptation from an appropriate shared response. The dialogue and waveform glyphs are illustrative.

ployment uses one utterance and the visible dialogue context; the paired recording and annotation cards are confined to learning and assessment.

Our contribution is the connection between acoustic representation, candidate-specific relational supervision, and bilateral evaluation. Matched learning comparisons test the response objective; constant-input and temporal-summary controls test the representation; quality and relation ablations test whether stronger differentiation is justified. This formulation makes an appropriate shared reply a valid outcome and makes factual errors visible alongside gains in suitability.

2 Related Work

From vocal perception to response generation. BLSP-Emo aligns speech and language for empathetic generation; TextPro-SLM and ES4R strengthen prosodic or affective input representations (Wang et al., 2024a; Cui et al., 2026; Gao et al., 2026). Prosody denotes patterns of pitch, timing, rhythm, and intensity accompanying the words. CogAudio-LLM combines affective reasoning and response alignment, while the Paralinguistics-Aware LLM (PALLM) jointly learns understanding and generation (Zhao et al., 2026; Chen et al., 2026a). These methods establish affect-sensitive generation as an existing research direction. ToneTrace uses a lightweight bottleneck and continuous prefix, both established mechanisms (Houlsby et al., 2019; Li & Liang, 2021); its role is to retain ordered acoustic information whose usefulness is tested by controlled response comparisons.

Relational supervision and evaluation. Counterfactual Relational Policy Optimization (CRPO) learns task-dependent response relations under paired video changes (Du et al., 2026). ARPO instantiates a related principle for uncertain vocal evidence: the difference is candidate-specific, activates only for reviewed contrastive families, and is constrained by ordinary reply quality. Its underlying policy update remains Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Existing evaluations separate lexical and acoustic emotion cues, conversational empathy, and tone-aware response generation (Chen et al., 2026b; Wang et al., 2026; Held et al., 2025). ToneFoil-Bench adds an explicit requirement that both acceptable replies be preferred in their respective contexts. This separates acoustic recognition, general empathy, and the narrower question of whether a particular response distinction is justified.

3 ToneFoil: Vocal Cues and Response Relations

Problem formulation. We study how vocal delivery changes the suitability of a response when the words and visible dialogue context are fixed. A family groups recordings with shared content u and visible context h. A reviewed comparison $F = (u, h, x_a, x_b, a_a, a_b, r)$ selects two recordings x_a, x_b, their acoustic evidence cards a_a, a_b, and a response relation r. Each card describes audible expression and its uncertainty. The response policy $\pi_\theta(y \mid x_i, h)$ receives only one recording

and the visible context; the paired recording, cards, and source labels are used for supervision and assessment.

A useful distinction concerns replies, not merely emotion labels. A *contrastive* family supports acceptable replies y_a, y_b such that $y_a \succ_a y_b$ and $y_b \succ_b y_a$: each reply is preferred in its own context. A *compatible* family admits an appropriate shared reply. An *uncertain* family lacks sufficient evidence for a stable preference. Acceptability requires factual correctness, relevance, and consistency with the assistant's role. These relations are conditional on the candidate replies and review rubric. They allow vocal evidence to change conversational emphasis without requiring every acoustic difference to produce a different response.

Construction. ToneFoil selects its initial recordings from Lexically-Identical, Multi-Emotion (LIME-440K) (Zhao et al., 2026). Construction proceeds from recording validity to content control, acoustic observation, and reply review. We remove duplicate waveforms and reject undecodable or unintelligible recordings. Content review preserves entities, numbers, negation, and discourse function, separating strictly identical wording from reviewed semantic equivalence. A label-blind observer then describes audible cues. Candidate replies are assessed for ordinary quality before their within-context preferences determine the relation. Source emotion labels do not determine reply preferences.

ToneFoil contains 100,000 distinct training recordings organized into 40,000 comparison families: 18,000 contrastive, 14,000 compatible, and 8,000 uncertain. The development set contains another 4,000 recordings. ToneFoil-Bench contains 4,000 additional recordings in 2,000 strictly same-word families, comprising 1,200 contrastive, 500 compatible, and 300 uncertain relations, with equal English and Chinese recording counts. The benchmark therefore evaluates a stricter content control than the semantically equivalent portion of training. Repeated replies or pair combinations do not increase the recording count.

Scope of the acoustic comparison. Matching words removes one source of variation in the response problem, but does not make two recordings identical in speaker, recording conditions, or delivery. The reviewed relation therefore concerns the suitability of specific replies under the available evidence. It does not label a latent mental state or establish that all responses must differ. For example, factual content can remain appropriate across a pair while a proposed reassurance requires stronger evidence. This distinction motivates separate relation labels and quality judgments: acoustic variation establishes an opportunity for adaptation, while reply review determines whether that adaptation is warranted.

Partitioning and assessment. A lexical parent groups identical or near-duplicate scripts; all its families remain in one partition. Waveform duplication, speaker identity, and source-corpus exposure are tracked separately, and previously exposed development examples are excluded from the test partition. Assessors receive anonymized audio, context, and counterbalanced replies. Quality and preference are judged separately, with explicit tie, both-unacceptable, and insufficient-evidence outcomes. These controls separate justified adaptation from label recognition or indiscriminate emotional differentiation. Source metadata, partition statistics, and license conditions appear in Appendix B.

4 VoxTact

VoxTact connects three operations: preserving vocal evidence, learning its relevance to reply suitability, and realizing the selected reply as speech. Figure 2 shows the complete architecture and learning stages. ToneTrace augments the original speech representation; Appropriateness-Relative Policy Optimization (ARPO) trains the response policy to use that information selectively. Neither component requires paired audio at deployment. Training uses supervised fine-tuning (SFT), followed by reinforcement learning (RL).

4.1 ToneTrace: Preserving Vocal Evidence

A speech representation optimized for lexical understanding need not expose all information relevant to response style. ToneTrace provides a short, ordered path from the speech encoder to

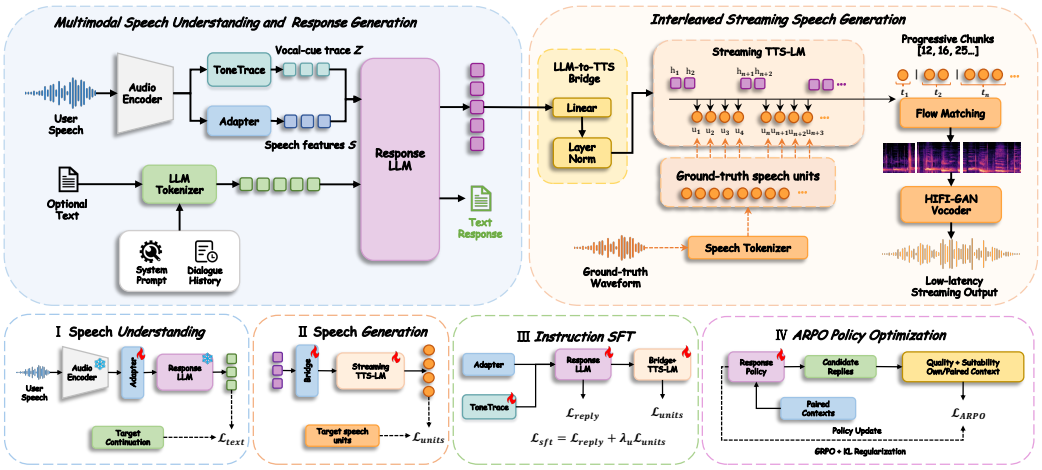

Figure 2: VoxTact architecture and learning objectives. ToneTrace adds vocal-cue summaries to the original speech representation. The response model's final-reply states condition streaming speech generation. The lower panels summarize speech understanding, speech generation, instruction SFT, and ARPO. Snowflakes and flames indicate frozen and trainable modules in the depicted updates. ToneTrace's reply-only adaptation occurs within Stage 3; the detailed update schedule is specified in Appendix A. Dashed arrows denote supervision or policy updates.

the response model while retaining the original representation for content understanding. Let $H = E(x) \in \mathbb{R}^{T \times d_e}$ be the valid encoder sequence and $S = A(H)$ the original speech-to-language representation. We average consecutive temporal regions into m summaries, then apply a shared two-layer projection:

$$B = \mathrm{Pool}_m(H), \qquad Z = \mathrm{GELU}(BW_1 + b_1)W_2 + b_2, \qquad S' = [Z; S]. \tag{1}$$

Here $B \in \mathbb{R}^{m \times d_e}$, $W_1 \in \mathbb{R}^{d_e \times d_b}$, and $W_2 \in \mathbb{R}^{d_b \times d_l}$; d_b is the bottleneck width and d_l is the response model's embedding dimension. GELU denotes the Gaussian Error Linear Unit (Hendrycks & Gimpel, 2016). We use eight summaries in temporal order. Unlike a single utterance average, they retain coarse changes over the recording, while shared projection weights avoid assigning a distinct semantic meaning to each position. The summaries can contain lexical, speaker, and affective information; their relevance is learned through reply supervision.

ToneTrace adds a fixed-length prefix while retaining the original encoder representation. Its eight summaries contribute eight input positions, independent of recording duration. Pooling still reads the valid encoder frames, so fixed prefix length does not imply constant total processing cost. The constant-input and repeated-global-summary controls match this prefix length and projection capacity.

ToneTrace is adapted within Stage 3 using the accepted final reply y^*:

$$\mathcal{L}_{\text{reply}} = -\sum_t \log \pi_\theta(y_t^* \mid x, h, y_{<t}^*). \tag{2}$$

Only ToneTrace's projection is updated during this adaptation; gradients pass through the fixed response model. The target is an appropriate reply, not an emotion label or an observation narrative. Subsequent policy learning holds the prepared input representation fixed, separating the availability of acoustic cues from the policy's use of them. Temporal pooling details and auxiliary-supervision comparisons appear in Appendices A and E.

4.2 ARPO: Learning Contextual Relevance

A candidate-specific comparison. An absolute suitability score can favor a generally helpful reply without showing why it fits one vocal context better. For candidate y, a frozen training scorer assigns ordinary quality $Q_i(y) \in [0, 1]$ and contextual suitability $J_i(y) \in [0, 1]$. Quality covers

correctness, relevance, and role consistency; suitability measures whether the reply's emphasis fits the audible evidence and visible context. ARPO holds the candidate fixed and exchanges its scoring context:

$$d_i(y) = J_i(y) - J_{\bar{i}}(y), \qquad c_i(y) = \mathrm{clip}(d_i(y)/\tau, -1, 1), \tag{3}$$

where $\bar{i}$ denotes the paired context and $\tau > 0$ sets the scale. A useful shared reply may score highly in both contexts and have a small difference. This behavior is desirable for compatible families.

Quality-constrained and relation-aware reward. A positive difference alone is insufficient: an exaggerated or unsupported claim can be highly context-specific. Let $g_i(y) = \mathbf{1}[Q_i(y) \geq q_0 \wedge J_i(y) \geq j_0]$ indicate quality and own-context acceptance, and let $b_F = 1$ only for reviewed contrastive families. The contextual term is

$$C_i(y) = b_F \big[g_i(y) \max(c_i(y), 0) + \min(c_i(y), 0) \big]. \tag{4}$$

Positive specificity receives credit only after acceptance, whereas a negative difference retains its penalty. Compatible and uncertain families have $b_F = 0$ and learn from ordinary quality and suitability. The total reward is

$$R_i(y) = \begin{cases} r_{\mathrm{fail}}, & \text{severe factual or role violation,} \\ w_Q Q_i(y) + w_J J_i(y) + \lambda_C C_i(y), & \text{otherwise,} \end{cases} \tag{5}$$

with nonnegative weights and a fixed failure reward r_{fail}. Thus, contextual specificity cannot compensate for a severe violation, and emotion labels alone cannot create contrastive supervision.

Policy optimization. ARPO uses Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each utterance, the policy generates a group of candidate replies; their rewards yield centered, scaled advantages $\widehat{A}[R]$. The loss combines the clipped GRPO policy term with Kullback–Leibler (KL) regularization toward the frozen supervised policy:

$$\mathcal{L}_{\mathrm{ARPO}} = \mathcal{L}_{\mathrm{GRPO}}\big(\widehat{A}[R]\big) + \beta\, \mathbb{E}\big[D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}}) \big]. \tag{6}$$

The difference from sample-level GRPO is the candidate-specific reward $C_i(y)$, not a different policy-gradient estimator. Advantages are normalized within each utterance, never across its paired contexts. A family-wide bonus shared by all candidates would disappear after centering. Appendix A.2 gives the token-level objective, and Appendix A specifies thresholds and weights.

The two constraints act at different levels. Quality acceptance is candidate-specific: a factual error blocks positive contextual credit even within a valid contrastive family. The relation mask is family-specific: compatible and uncertain families retain the base reward without an incentive to exaggerate their difference. Within an accepted contrastive family, candidates can still receive different contextual rewards. ARPO therefore learns which responses use the available cues appropriately, while preserving ordinary response learning when the review does not support a contrast. Setting $\lambda_C = 0$ recovers the base-reward GRPO comparison.

The paired recording changes how a candidate is scored, not what the candidate-generating policy observes. During supervised adaptation and policy learning, the student receives its own recording and ordinary dialogue context. Evidence cards inform supervision and review; they are not additional response inputs. This separation prevents a privileged description of the paired condition from becoming an inference requirement.

Training and assessment answer distinct questions. Training scores *one reply in two contexts*; final assessment compares *two replies in one context*, then requires appropriate preferences on both sides. The latter additionally enforces quality for both generated replies. This distinction prevents a high training difference from being treated as bilateral behavioral success; the scoring matrix is given in Appendix C.1.

4.3 Supervised Preparation and Spoken Output

Stage 1 learns speech-conditioned text continuation with $\mathcal{L}_{\mathrm{text}}$; Stage 2 learns reference-reply-conditioned speech-unit prediction with $\mathcal{L}_{\mathrm{units}}$. Stage 3 jointly supervises replies and speech through $\mathcal{L}_{\mathrm{SFT}} = \mathcal{L}_{\mathrm{reply}} + \lambda_u \mathcal{L}_{\mathrm{units}}$ and includes ToneTrace's reply-only adaptation. Stage 4 then optimizes

the response policy with ARPO. Dataset mixtures, freezing schedules, and the supervised loss expansions are in Appendix A.

The final reply's states condition a speech-unit language model through a projection and normalization bridge. An acoustic decoder and vocoder convert the units into waveform output. This text-to-speech (TTS) path receives neither observation cards nor scorer explanations. It remains fixed during policy optimization, while the response policy changes; speech naturalness and text–audio consistency are therefore evaluated alongside reply suitability.

5 EXPERIMENTS

We evaluate whether VoxTact makes better use of vocal evidence, whether the benefit transfers beyond ToneFoil, and whether it preserves reply and speech quality. Matched learning comparisons test ARPO, factorial controls separate its contribution from ToneTrace, and targeted ablations examine the information and constraints each component supplies.

5.1 EVALUATION SETUP

Model instantiation. VoxTact uses a Qwen3.5-9B response model and the Qwen2-Audio speech encoder (Qwen Team, 2026a; Chu et al., 2024). A 0.8B speech-unit language model conditions on final-reply states, with speech units and waveform reconstruction following GLM-4-Voice (Qwen Team, 2026b; Zeng et al., 2024). The compared response-learning methods share this prepared system. The supervised data mixtures and parameter-update schedule are specified in Appendix A.

ToneFoil-Bench. Bilateral response success (BRS) measures the central behavioral distinction. For each contrastive family, both generated replies must pass quality review and each must be preferred in its own context. With quality decision V_i and within-context preference P_i, we compute

$$\mathrm{BRS} = \frac{1}{|\mathcal{F}_c|} \sum_{F \in \mathcal{F}_c} \prod_{i \in \{a,b\}} \mathbf{1}[V_i(y_i)] \mathbf{1}[P_i(y_i \succ y_{\bar{i}})]. \tag{7}$$

Ties and two unacceptable replies are not strict successes. We average three repetitions on the same 1,200 contrastive families. Quality acceptance (Q) and unsupported-assertion rate (UA) cover all 4,000 benchmark replies, including compatible and uncertain relations. Pair emotion accuracy (PEA), reported in the appendix, requires both independently elicited emotion predictions to be correct. BRS therefore tests response selection rather than emotion recognition alone.

The metrics deliberately cover different family sets. Strict BRS is defined on contrastive families, where opposite within-context preferences are supported. Applying that success criterion to compatible families would penalize an appropriate shared answer. Conversely, restricting all evaluation to contrastive families would hide errors introduced on neutral or ambiguous inputs. Q and UA therefore cover every response relation. PEA is computed on the 1,600 families with supported discrete emotion labels. These denominators are kept separate: recognizing both source emotions does not establish that the generated replies satisfy bilateral preference or ordinary quality.

Interpreting quality and specificity. Quality acceptance is a final binary decision about a generated reply, distinct from the continuous training score $Q_i(y)$. Unsupported assertions include claims about facts, the speaker's mental state, or actions that the assistant has not performed. A reply can therefore acknowledge an audible hesitation without asserting its cause, or offer reassurance without claiming certainty about the speaker's feelings. This distinction is central to the evaluation: greater sensitivity to vocal delivery is useful only when the resulting response remains justified.

The final assessor compares replies under a single recording with system identities hidden and reply order counterbalanced. An appropriate shared answer can pass quality review in both contexts while failing strict BRS, because it does not establish a bilateral preference. Conversely, a strongly differentiated pair fails BRS if either reply is unacceptable. The metric thus rewards a conjunction of quality and contextual preference, rather than textual diversity. Reporting Q and UA alongside BRS exposes whether a gain comes with a broader cost on compatible or uncertain families.

Table 1: Matched response learning and external transfer. BRS, Q, and UA are ToneFoil-Bench percentages. LISTEN reports mean conflict accuracy across audio-only and text-plus-audio inputs; HumDial reports conflict accuracy and Qwen3-Omni textual empathy (1–5). Each column uses a common assessment protocol across methods. Bold/underline mark best/second.

	ToneFoil-Bench			LISTEN	HumDial	
Method	BRS ↑	Q ↑	UA ↓	Conflict ↑	Conflict ↑	Text ↑
VoxTact	**63.34**	**94.10**	**2.90**	**72.10**	**75.50**	**4.12**
Paired SFT	39.50	92.80	4.20	61.80	65.50	3.66
DPO (Rafailov et al., 2023)	44.00	<u>93.30</u>	<u>3.55</u>	65.30	69.00	3.86
Sample GRPO (Shao et al., 2024)	<u>49.94</u>	93.10	3.75	<u>67.40</u>	<u>71.00</u>	<u>3.92</u>

Learning comparisons. Paired supervised fine-tuning (SFT), Direct Preference Optimization (DPO) (Rafailov et al., 2023), sample-level GRPO, and ARPO share the prepared ToneTrace representation and response-learning families. Candidate exposure, response-token budget, and optimization opportunities are matched; scorer calls are accounted for separately. The factorial experiment additionally disables ToneTrace. All four factorial conditions undergo matched policy learning, including F00. The unadapted base in the voice-assistant table is a separate reference.

External tasks and assessors. VoiceBench evaluates open responses and spoken knowledge (Chen et al., 2026c); the Understanding, Reasoning, and Oral conversation benchmark (URO-Bench) supplies UnderEmotion (Yan et al., 2025). The Lexical vs. Acoustic Speech Test for Emotion in Narratives (LISTEN) tests competing lexical and acoustic cues (Chen et al., 2026b). The Human-Recorded Multi-Turn Emotional Intelligence Benchmark (HumDial-EIBench) evaluates conflict recognition and response empathy (Wang et al., 2026). LISTEN scores average its native audio-only and text-plus-audio conditions. HumDial textual empathy is scored by Qwen3-Omni (Qwen Team, 2025); generated vocal congruence and naturalness receive separate human ratings. These text and speech assessments are not interchangeable.

The voice-assistant comparison includes speech models, an automatic speech recognition (ASR) cascade, and multimodal assistants evaluated through their speech/text interface. No system receives gold emotion labels or a privileged transcript; the cascade transcribes its own input. Subjective scores in Table 2 use the same DeepSeek-V4 judge and rubric (DeepSeek-AI, 2026). Detailed benchmark protocols, speech results, and EchoMind evaluations (Zhou et al., 2026) appear in Appendix D. Rankings describe comparable point estimates, not statistical significance.

5.2 Response Suitability and External Transfer

ARPO improves BRS from 49.94% with sample GRPO to 63.34%, a 13.40-point gain under matched response learning (Table 1). Quality acceptance increases to 94.10%, while unsupported assertions decline to 2.90%. Paired SFT reaches 39.50% and DPO 44.00%. Exposure to appropriate replies is therefore useful but does not match learning which candidate better fits each vocal context.

The same comparison transfers to external conflict and empathy tasks. VoxTact obtains 72.10% LISTEN conflict accuracy, exceeding sample GRPO by 4.70 points. HumDial conflict accuracy increases from 71.00% to 75.50%, and textual empathy rises from 3.92 to 4.12/5. Conflict accuracy tests whether acoustic cues remain usable when the words suggest a competing interpretation; textual empathy tests whether the response acknowledges and addresses the speaker's situation. Together, these outcomes extend the controlled bilateral result to different tasks without treating emotion recognition as sufficient evidence of empathy.

The detailed LISTEN conditions further locate the transfer gain. Relative to sample GRPO, neutral-text accuracy rises from 75.80 to 77.60 and emotion-matched accuracy from 82.60 to 84.30, compared with the larger 4.70-point gain under conflict. The strongest observed improvement thus occurs when lexical and acoustic cues compete. On HumDial, the human vocal-congruence rating changes only from 3.77 to 3.81, while naturalness remains 3.94. Improved textual empathy is therefore accompanied by a much smaller change in perceived vocal expression. These results support better response selection, with a separate empirical boundary on expressive speech generation.

Table 2: Study reevaluation of voice assistants. A/C/W are AlpacaEval/CommonEval/WildVoice ratings (1–5); UE denotes UnderEmotion (/100). All subjective columns use the common DeepSeek-V4 rubric. An em dash denotes an unreported metric. Bold/underline mark best/second within the comparable reevaluation group.

Type	Model	A ↑	C ↑	W ↑	UE-en ↑	UE-zh ↑
Speech	Qwen2-Audio (Chu et al., 2024)	3.86	3.38	2.97	—	—
Cascade	ASR–LLM–TTS cascade	4.49	4.05	**4.08**	58.54	90.38
Dialogue	GLM-4-Voice (Zeng et al., 2024)	3.97	3.37	2.98	68.61	81.85
	Freeze-Omni (Wang et al., 2024b)	4.10	3.51	3.15	64.53	77.72
	LLaMA-Omni2 (Fang et al., 2025)	3.69	3.34	2.99	47.01	74.18
	OpenS2S (Wang et al., 2025)	4.01	3.65	3.51	57.37	70.38
	OSUM-EChat (Geng et al., 2025)	4.14	3.59	2.93	67.78	**92.96**
	Step-Audio 2 mini (Wu et al., 2025)	3.69	3.16	3.13	58.69	75.70
Omni	Qwen2.5-Omni (Xu et al., 2025)	4.28	3.85	3.73	61.61	84.30
	MiniCPM-o 2.6 (MiniCPM-o Team, OpenBMB, 2025)	4.41	4.04	3.84	<u>70.66</u>	88.86
	Baichuan-Omni-1.5 (Li et al., 2025)	**4.60**	4.07	3.90	48.32	64.30
Proposed	VoxTact	<u>4.51</u>	**4.21**	<u>3.92</u>	**71.74**	<u>91.82</u>
	w/o ToneTrace	4.43	<u>4.19</u>	3.84	63.26	83.75
	w/o ARPO	4.40	4.18	3.82	60.55	78.48
	w/o both	4.37	4.17	3.78	54.59	76.96

5.3 Voice-Assistant Quality

VoxTact attains UnderEmotion scores of 71.74 in English and 91.82 in Chinese, compared with 54.59 and 76.96 for the unadapted base. Removing ToneTrace reduces these scores by 8.48 and 8.07 points; removing ARPO reduces them by 11.19 and 13.34 points. Open-response ratings also improve, from 4.37/4.17/3.78 to 4.51/4.21/3.92 on AlpacaEval/CommonEval/WildVoice. The gains are largest on affect-sensitive tasks, while general response quality remains competitive. The two UnderEmotion language subsets contain different utterances and are interpreted separately.

Speech evaluation compares generated audio with the model's own reply. VoxTact obtains English/Chinese automatic naturalness scores of 4.46/3.75 and word/character error rates of 1.89%/1.23%. These results assess naturalness and content fidelity; human vocal-empathy results are reported separately in Appendix D. On NVIDIA RTX A6000 hardware under the common timing protocol, the first playable audio arrives in 1.076 s (Appendix G).

The speech changes are small and not uniformly favorable. Relative to the base, English word error rate decreases from 1.95% to 1.89%, while Chinese character error rate increases from 1.21% to 1.23%. Automatic naturalness rises from 4.35/3.65 to 4.46/3.75. A fixed speech generator can receive different reply states after policy learning, so parameter freezing alone is not a capability-retention result. Reporting both transcription agreement and naturalness makes this distinction observable and avoids inferring preservation solely from which parameters were updated.

5.4 Component Contributions

Representation and objective. Table 3 combines the component controls with the trained 2×2 comparison. ToneTrace adds 12.53 points under sample GRPO and 11.36 under ARPO; ARPO adds 14.57 points without ToneTrace and 13.40 with it. The main effects are +11.95 and +13.99 points, with a -1.17-point interaction. Both components help, with a slightly subadditive joint effect. F10 and F11 are the sample-GRPO and VoxTact conditions in Table 1.

The constant-input control retains ToneTrace's capacity but removes recording-dependent information; its BRS falls to 50.25%. Repeating one global summary preserves prefix length but removes temporal variation, yielding 56.00%. Thus, neither additional capacity nor a repeated global representation accounts for the full result. Shuffling family assignments yields 49.50%, and scoring with

Table 3: Component evidence. (a) Each ablation changes one representation or reward component; all scores are percentages. (b) BRS for the trained 2×2 design: rows enable ToneTrace, columns change the reward. F00 is a trained, matched-budget control. Rankings are computed separately for panels (a) and (b).

(a) Representation and reward

Condition	BRS ↑	Q ↑	UA ↓
VoxTact	**63.34**	**94.10**	**2.90**
Constant input	50.25	93.20	3.65
Global summary	56.00	<u>93.70</u>	<u>3.20</u>
Shuffled families	49.50	93.05	4.05
No quality gate	57.00	81.00	13.85
All contrastive	<u>57.50</u>	88.20	8.70
Wrong scoring context	51.00	92.90	4.50

(b) Trained factorial controls

ToneTrace	GRPO	ARPO
Off	37.41	<u>51.98</u>
On	49.94	**63.34**

Rows off/on; columns GRPO/ARPO define F00/F01/F10/F11.

Main effects: ToneTrace +11.95 pp; ARPO +13.99 pp. Interaction: -1.17 pp.

mismatched contexts yields 51.00%, supporting the importance of the reviewed relation and its use in the reward.

The auxiliary-target comparison also separates recognition from response learning. Adding absolute and difference-based affect targets yields 85.00% PEA, above the full model's 82.50%, but achieves 60.25% BRS rather than 63.34%. These measures use their respective labeled and contrastive family sets. The result does not make affect recognition unimportant; it shows that the auxiliary target most effective for paired recognition is not the strongest objective for bilateral reply suitability in this comparison. Accepted-reply supervision remains the ToneTrace objective used by the full model.

Specificity requires quality constraints. Without quality gating, BRS falls to 57.00%, quality acceptance drops from 94.10% to 81.00%, and unsupported assertions rise from 2.90% to 13.85%. Treating every family as contrastive similarly increases unsupported assertions to 8.70%. The constraints improve bilateral suitability as well as factual restraint: forcing stronger differentiation is not a substitute for selecting an appropriate reply. Table 3 places these outcomes beside the representation controls; Appendix E provides the auxiliary-target and bilateral-DPO comparisons and a joint visualization of BRS, quality, and unsupported assertions.

6 Discussion and Limitations

The findings distinguish preserved information from appropriate use. The real-input control improves over an otherwise identical constant-input adapter, while ARPO improves bilateral suitability over sample-level GRPO. Removing quality constraints increases unsupported assertions even when replies remain differentiated. Together, these comparisons support selective acoustic conditioning under explicit reply constraints.

The interpretation has limits. Ordered encoder summaries are not a pure affect representation, and paired recordings do not isolate a speaker's latent psychological state. Shared source corpora, uncertain speaker identities, and unknown foundation-model exposure limit generalization claims. Automatic observations and judgments can retain correlated errors despite role separation and order balancing. Aggregate scores do not establish human consensus, clinical validity, or statistical significance across training seeds. Speech naturalness and transcription consistency also do not measure whether the generated voice expresses appropriate empathy.

7 Conclusion

VoxTact connects vocal evidence to response suitability through ToneTrace, ARPO, and ToneFoil. ToneTrace preserves ordered acoustic information; ARPO makes its relevance candidate-specific and quality-constrained; ToneFoil-Bench tests whether both replies are justified under fixed words. The resulting formulation values an appropriate shared reply alongside supported adaptation.

AI USE STATEMENT

Generative artificial intelligence tools assisted with language polishing and translation, retrieval and discovery of related literature, and drafting manuscript sections and figures. Model-based observation and scoring are part of the methodology described in Sections 3–5 and Appendix F. The authors reviewed the AI-assisted material and take responsibility for the text, citations, scientific claims, and reported results.

ETHICS STATEMENT

Vocal annotations describe perceived expression in context, not diagnoses or privileged knowledge of a speaker's intentions. The response objective penalizes unsupported factual claims, role reversals, and false commitments. Reuse and redistribution of speech remain governed by the original sources' permissions; secondary annotation does not establish new ownership or consent. Human empathy ratings and automated response assessments are distinguished throughout the evaluation. The conclusions concern response suitability within the evaluated settings and do not justify sensitive speaker profiling.

REPRODUCIBILITY STATEMENT

Appendix A specifies the model, supervised data mixtures, and optimization objective. Appendix B describes data selection and partitioning; Appendices C–D define the metrics and evaluation protocols. Ablations, prompt templates, and timing procedures are given in Appendices E–G. The accompanying source materials contain the numerical results underlying the tables and figures and a provenance record that distinguishes evaluation protocols and model identities.

REFERENCES

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. Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020. URL https://aclanthology.org/2020.lrec-1.520/.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio. arXiv preprint, 2021. URL https://arxiv.org/abs/2106.06909.

Jingxiang Chen, Minseok Kim, Seong-Gyun Leem, Yin Huang, Rashi Rungta, Zhicheng Ouyang, Haibin Wu, Surya Teja Appini, Ankur Bansal, Yang Bai, Yue Liu, Florian Metze, Ahmed A Aly, Anuj Kumar, Ariya Rastrow, and Zhaojiang Lin. Aligning Paralinguistic Understanding and Generation in Speech LLMs via Multi-Task Reinforcement Learning. arXiv preprint, 2026a. URL https://arxiv.org/abs/2603.15981v1.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. Do Audio LLMs Really LISTEN, or Just Transcribe? Measuring Lexical vs. Acoustic Emotion Cues Reliance. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5848–5877, 2026b. doi: 10.18653/v1/2026.eacl-long.274. URL https://aclanthology.org/2026.eacl-long.274/.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-Based Voice Assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398, 2026c. doi: 10.1162/tacl.a.628. URL https://aclanthology.org/2026.tacl-1.18/.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio Technical Report. arXiv preprint, 2024. URL https://arxiv.org/abs/2407.10759.

Wenqian Cui, Xiao-Hui Li, Daxin Tan, Qiyong Zheng, and Irwin King. Minimizing Modality Gap from the Input Side: Your Speech LLM Can Be a Prosody-Aware Text LLM. arXiv preprint, 2026. URL https://arxiv.org/abs/2605.05927v2.

DeepSeek-AI. DeepSeek-V4 Preview Release. Official model release, April 24, 2026. URL https://api-docs.deepseek.com/news/news260424/.

Dazhao Du, Jian Liu, Jialong Qin, Tao Han, Bohai Gu, Fangqi Zhu, Yujia Zhang, Eric Liu, Xi Chen, and Song Guo. Learning Spatiotemporal Sensitivity in Video LLMs via Counterfactual Reinforcement Learning. arXiv preprint, 2026. URL https://arxiv.org/abs/2605.21988v1.

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, 2024. URL https://arxiv.org/abs/2412.10117.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. LLaMA-Omni: Seamless Speech Interaction with Large Language Models. arXiv preprint, 2024. URL https://arxiv.org/abs/2409.06666.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. LLaMA-Omni2: LLM-based Real-time Spoken Chatbot with Autoregressive Streaming Speech Synthesis. arXiv preprint, 2025. URL https://arxiv.org/abs/2505.02625.

Zhuoyue Gao, Xiaohui Wang, Xiaocui Yang, Wen Zhang, Daling Wang, Shi Feng, and Yifei Zhang. ES4R: Speech Encoding Based on Prepositive Affective Modeling for Empathetic Response Generation. arXiv preprint, 2026. URL https://arxiv.org/abs/2601.16225v1.

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, 2025. URL https://arxiv.org/abs/2508.09600.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An Extensive, Multilingual, and Diverse Speech Dataset for Large-Scale Speech Generation. arXiv preprint, 2024. URL https://arxiv.org/abs/2407.05361.

Will Held, Michael J. Ryan, Aditya Shrivastava, Ali Sartaz Khan, Caleb Ziems, Ella Li, Martijn Bartelds, Michael Sun, Tan Li, Woody Gan, and Diyi Yang. CAVA: Comprehensive Assessment of Voice Assistants . Official benchmark website, 2025. URL https://talkarena.org/cava.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). arXiv preprint, 2016. URL https://arxiv.org/abs/1606.08415.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP. In *International Conference on Machine Learning*, pp. 2790–2799, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

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. arXiv preprint, 2021. URL https://arxiv.org/abs/2106.09685.

Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi Ma, Kegeng Wu, Jia li, Aiyuan Yang, Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin Huang, Kun Li, Wenjing Luo, Yifei Duan, Lingling Zhu, Ran Xiao, Zhe Su, Jiani Pu, Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yujing Qiao, Lei Zhang, Yanjun Shen, Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao, Hao Liang, Youquan Li, Yanzhao Qin, Linzhuang Sun, Jianhua Xu, Haoze Sun, Mingan Lin, Zenan Zhou, and Weipeng Chen. Baichuan-Omni-1.5 Technical Report. arXiv preprint, 2025. URL https://arxiv.org/abs/2501.15368.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004. URL https://aclanthology.org/W04-1013/.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards Emotional Support Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3469–3483, 2021. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269/.

MiniCPM-o Team, OpenBMB. MiniCPM-o 2.6: A GPT-4o Level MLLM for Vision, Speech, and Multimodal Live Streaming on Your Phone. Official technical release, January 24, 2025. URL https://openbmb.vercel.app/minicpm-o-2-6-en.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. Expresso: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis. In *Proc. Interspeech 2023*, pp. 4823–4827, 2023. doi: 10.21437/Interspeech.2023-1905. URL https://www.isca-archive.org/interspeech_2023/nguyen23_interspeech.html.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. IEEE International Conference on Acoustics, Speech and Signal Processing, 2015. URL https://www.openslr.org/12/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. arXiv preprint, 2024. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3-Omni Technical Report, 2025. URL https://arxiv.org/abs/2509.17765v1.

Qwen Team. Qwen3.5-9B Model Card. Official model card, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Qwen Team. Qwen3.5-0.8B Model Card. Official model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.5-0.8B.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, 2019. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534/.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Proc. Interspeech 2022*, pp. 4521–4525, 2022. doi: 10.21437/Interspeech.2022-439. URL https://www.isca-archive.org/interspeech_2022/saeki22c_interspeech.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv preprint, 2024. URL https://arxiv.org/abs/2402.03300v3.

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, 2024a. 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, 2025. doi: 10.18653/v1/2025.emnlp-demos.70. URL https://aclanthology.org/2025.emnlp-demos.70/.

Shuiyuan Wang, Zhixian Zhao, Hongfei Xue, Chengyou Wang, Shuai Wang, Hui Bu, Xin Xu, and Lei Xie. HumDial-EIBench: A Human-Recorded Multi-Turn Emotional Intelligence Benchmark for Audio Language Models. arXiv preprint, 2026. URL https://arxiv.org/abs/2604.11594v2.

Xiong Wang, Yangze Li, Chaoyou Fu, 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. arXiv preprint, 2024b. URL https://arxiv.org/abs/2411.00774.

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. arXiv preprint, 2025. URL https://arxiv.org/abs/2507.16632.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni Technical Report. arXiv preprint, 2025. URL https://arxiv.org/abs/2503.20215.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. URO-Bench: Towards Comprehensive Evaluation for End-to-End Spoken Dialogue Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17211–17242, 2025. doi: 10.18653/v1/2025.findings-emnlp.933. URL https://aclanthology.org/2025.findings-emnlp.933/.

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, 2024. URL https://arxiv.org/abs/2412.02612.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. WenetSpeech: A 10000+ Hours Multi-domain Mandarin Corpus for Speech Recognition. arXiv preprint, 2021. URL https://arxiv.org/abs/2110.03370.

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. arXiv preprint, 2023. URL https://arxiv.org/abs/2305.11000.

Zhixian Zhao, Shuiyuan Wang, Wenjie Tian, Jingbin Hu, Ziyu Zhang, and Lei Xie. Beyond Semantic Dominance: Cognitive Affective Reasoning and Empathetic Response Alignment in Audio Language Models. arXiv preprint, 2026. URL https://arxiv.org/abs/2606.06940v2.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional Voice Conversion: Theory, Databases and ESD. arXiv preprint, 2021. URL https://arxiv.org/abs/2105.14762.

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