# SPEECHANCHOR: ENHANCING SPEECH LARGE LANGUAGE MODELS FOR CONSISTENT MULTI-TURN SPOKEN DIALOGUE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

End-to-end Speech Large Language Models (SLMs) have shown strong potential for natural voice interaction by directly modeling speech-to-speech communication. However, practical voice interaction is inherently multi-turn, requiring SLMs to maintain semantic coherence, adapt to evolving paralinguistic cues, and preserve speaker style throughout the conversation. Existing SLMs remain limited in multi-turn dialogue because most training corpora are dominated by single-turn conversations, and existing architectures mainly rely on the LLM backbone to track information across dialogue history. To address these challenges, we propose **SpeechAnchor**, a multi-turn consistency enhancement framework for SLMs. SpeechAnchor combines a lightweight module that explicitly models cross-turn speech history for the speech decoder with a consistency-aware data pipeline that constructs semantically coherent, speaker-consistent, and paralinguistically rich multi-turn speech data. Experiments on three multi-turn and four single-turn speech benchmarks show consistent gains in multi-turn consistency without compromising general speech performance. These results establish SpeechAnchor as a practical framework for advancing multi-turn SLMs and provide a unified perspective to guide future research on multi-turn consistency.[1]

## 1 INTRODUCTION

End-to-end Speech Large Language Models (SLMs) have advanced speech-to-speech interaction by directly modeling speech input and output, enabling more natural responses while preserving rich acoustic information (Zhang et al., 2023a; Fang et al., 2024; Défossez et al., 2024; Zeng et al., 2024; Xie & Wu, 2024; Chen et al., 2025). Since real-world voice interaction is inherently multi-turn, SLMs must maintain consistency across conversational turns to support reliable and natural interaction.

We believe that ideal multi-turn speech dialogue demands consistency across three dimensions. First, **semantic consistency** requires that responses align with the overall dialogue context, remain free of self-contradiction, and continuously retain and leverage information from previous turns (Kim et al., 2025; Yan et al., 2025; Ma et al., 2025; Gosai et al., 2025). Second, **paralinguistic consistency** necessitates accurate recognition of paralinguistic cues such as emotion, intonation, and prosody in the current turn, as well as their evolution across turns, enabling the model to adapt its responses accordingly (Liu et al., 2025; Du et al., 2025a). Third, **speaker style consistency** mandates that attributes including speaker identity, volume, and speaking rate remain stable throughout the interaction (Lin et al., 2025). Together, these three forms of consistency determine whether an SLM can deliver a truly natural and trustworthy conversational experience.

Despite its importance for natural spoken interaction, multi-turn consistency in existing SLMs remains constrained by two key bottlenecks. On the architecture side, existing approaches mainly rely on the LLM backbone to model dialogue history, such that during speech generation, each speech position passed to the speech decoder is conditioned only on its corresponding LLM hidden state, forcing the LLM backbone to simultaneously track and integrate semantic content, paralinguistic cues, and speaker-style information from dialogue history, which imposes a heavy modeling burden (Chen et al., 2025; Cheng et al., 2025). On the data side, most available speech-to-speech training

[1] Code are available in supplementary materials.



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.

corpora are dominated by single-turn conversation, while existing public multi-turn datasets are often biased toward daily conversational scenarios, with short dialogue lengths and highly colloquial expressions (Busso et al., 2008; Lee et al., 2023; Poria et al., 2019).

In this paper, we propose **SpeechAnchor**, a multi-turn consistency enhancement framework for SLMs. On the **architecture** side, we introduce the SpeechAnchor module, implemented with Gated DeltaNet (Yang et al., 2024), to explicitly model speech-modality dialogue history and inject the resulting representation into the speech decoder. Through gated recurrent state updates, the module maintains compact long-term memory over historical speech states and retrieves relevant cross-turn information conditioned on the current speech hidden state, providing the speech decoder with a dedicated pathway for history modeling. This decoupled design reduces the modeling burden on the LLM backbone and enables more effective cross-turn context modeling. On the **data** side, we rewrite high-quality multi-turn text dialogue into speech-friendly form, then synthesize semantically coherent dialogues with consistent speaker identities. We further mine paralinguistically sensitive cases from textual dialogues and augment them with context-consistent cue annotations and adaptive responses to construct paralinguistically rich speech data.

We train SpeechAnchor with a two-stage strategy that first establishes semantic consistency and then enhances paralinguistic sensitivity using paired neutral and paralinguistic-augmented dialogues. Across three multi-turn and four single-turn speech benchmarks, SpeechAnchor improves multi-turn consistency while preserving strong general speech performance. Our main contributions are as follows:

- **SpeechAnchor** introduces an additional decoder-side pathway that explicitly models historical speech and directly conditions the speech decoder on a recurrent cross-turn speech state, providing a new architectural direction for multi-turn speech modeling.
- We develop a **data construction framework** that builds semantically coherent, speaker-consistent multi-turn speech dialogue and is the first to mine and annotate paralinguistically sensitive cases from text-only dialogues with context-consistent cues and response criteria.
- We frame multi-turn consistency in end-to-end SLMs across semantic, paralinguistic, and speaker-style dimensions, providing a unified perspective for future research in multi-turn spoken dialogue.

## 2 Related Work

**Speech Large Language Model.** End-to-end Speech Large Language Models (SLMs) enable direct speech-to-speech interaction without cascaded ASR and TTS modules (Zhang et al., 2023a; Fang et al., 2024; Défossez et al., 2024; Zeng et al., 2024; Xie & Wu, 2024; Chen et al., 2025). A typical SLM consists of a speech encoder, an LLM backbone, and a speech decoder. The speech encoder often adopts pretrained ASR models (Radford et al., 2023; Hsu et al., 2021) or specialized encoders for semantic and acoustic representations (Tang et al., 2023; Chu et al., 2024; Das et al., 2024). The LLM backbone processes speech representations and generates responses, either by adapting pretrained text LLMs with modality-bridging modules (Fang et al., 2024; Xie & Wu, 2024; Chen et al., 2025) or by pretraining on interleaved speech-text data (Défossez et al., 2024; Zeng et al., 2024). The speech decoder typically relies on neural audio codecs with residual vector quantization (Zeghidour et al., 2021; Défossez et al., 2022; Zhang et al., 2023b) to generate speech waveforms from discrete tokens.

**Multi-turn Enhancement of SLMs.** Multi-turn SLMs remain challenging, with prior work mainly focusing on data construction and historical context modeling. On the data side, SLAM-Omni converts text-based multi-turn dialogues into speech via TTS and constrains utterance length to better match daily conversations (Chen et al., 2025), while Audio Flamingo 3 retrieves acoustically diverse audio and prompts LLMs to generate dialogues conditioned on transcriptions and speech metadata (Goel et al., 2025). On the modeling side, SLAM-Omni compresses historical speech into textual prompts (Chen et al., 2025), OmniChat dynamically fuses semantic, emotional, and acoustic features (Cheng et al., 2025), while InteractiveOmni and Audio Flamingo 3 treat multi-turn capability as a dedicated alignment stage (Tong et al., 2025; Goel et al., 2025). Our work extends this line by constructing paralinguistically rich multi-turn speech data and introducing the SpeechAnchor module to explicitly model speech-modality dialogue history.

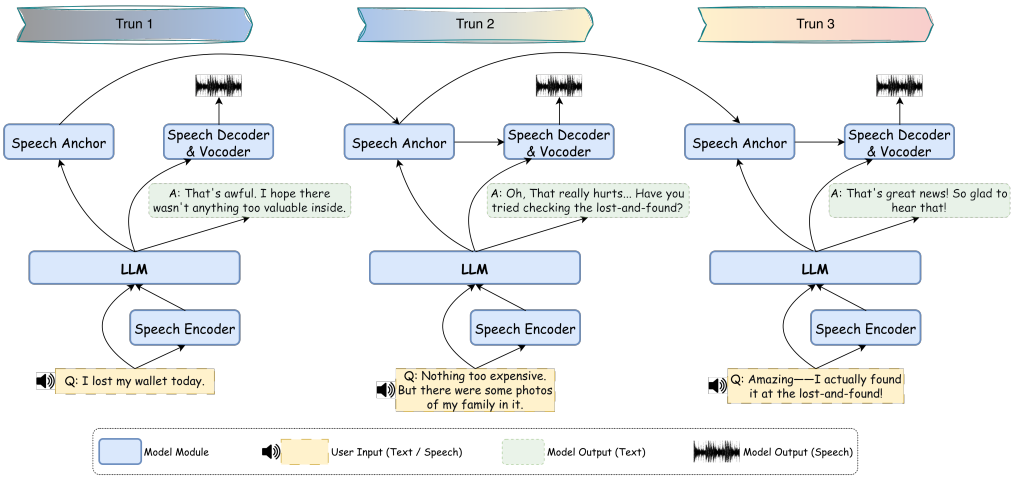



Figure 1: Overview of the SpeechAnchor architecture.

# 3 METHOD

This section introduces SpeechAnchor, a framework for improving multi-turn consistency in SLMs. SpeechAnchor combines a data synthesis pipeline that constructs semantically coherent and paralinguistically rich speech dialogues with an anchor module that injects historical speech-modality context into the speech decoder. The model is trained end-to-end with a two-stage strategy aligned with the synthesized data. The overall framework is illustrated in Figure 1.

## 3.1 DATA SYNTHESIS PIPELINE

Existing public multi-turn dialogue datasets (Koudounas et al., 2025; Busso et al., 2008; Park et al., 2024; Lee et al., 2023; Poria et al., 2019) are valuable resources for spoken dialogue modeling. However, they are often biased toward daily conversational scenarios, with short dialogue lengths, simple responses, and highly colloquial expressions. Directly training on such data may weaken the model's reasoning and instruction-following abilities in multi-turn interaction. To build high-quality multi-turn speech dialogue data, we design a two-stage synthesis pipeline that first converts rich textual dialogues into semantically coherent speech, and then augments them with diverse paralinguistic variations.

### 3.1.1 SEMANTIC CONSISTENCY DATA

We start from a high-quality multi-turn textual dialogue corpus. To ensure suitability for speech synthesis, we filter out samples containing non-verbalizable content, such as code and complex math problems. We then rewrite the responses in each multi-turn dialogue, such as converting Markdown-style answers into more speech-friendly expressions, using the prompt provided in Appendix C. The curated dialogues form the foundation set $D_{\text{sem}}^{\text{text}} = \{\mathcal{T}_j\}$, where each $\mathcal{T}_j = \{(q_i, r_i)\}_{i=1}^{n}$ denotes an $n$-turn dialogue.

To synthesize realistic two-party conversations, we assign each dialogue $\mathcal{T}_j$ a consistent user speaker $\text{ID}_k^{\text{user}}$ and assistant speaker $\text{ID}_k^{\text{asst}}$. User questions and assistant responses are synthesized with Instruct TTS model:

$$q_i^{\text{sem}} = \text{TTS}_{\text{inst}}\big(q_i; \text{spk} = \text{ID}_k^{\text{user}}\big), \tag{1}$$

$$r_i^{\text{sem}} = \text{TTS}_{\text{inst}}\big(r_i; \text{spk} = \text{ID}_k^{\text{asst}}\big), \quad \forall i \in \{1, \ldots, n\}. \tag{2}$$

The resulting dataset provides supervision for learning multi-turn semantic coherence while preserving stable role-specific speaker identities across the dialogue.

#### 3.1.2 PARALINGUISTIC CONSISTENCY DATA

Text-only dialogues usually omit paralinguistic cues that are critical in spoken interaction, such as emotion, stress, intonation, speaking rate, and pitch. To enrich the user queries with such cues while preserving high-quality responses, we construct an augmented dataset from $D_{\text{sem}}^{\text{text}}$ through filtering, annotation, response rewriting, and speech synthesis, using the prompts provided in Appendix C.

**Step 1: Contextual Filtering.** Starting from the filtered dialogues $D_{\text{sem}}^{\text{text}}$, we use LLM as a context-aware judge to select conversations where adding or modifying paralinguistic information would lead to clearly different assistant responses. This yields $D_{\text{para}}^{\text{text}}$. The filtering focuses on dialogues whose user intent or emotional state can be meaningfully affected by acoustic cues.

**Step 2: Paralinguistic Annotation.** For each selected dialogue, we prompt the LLM to assign concise and contextually aligned paralinguistic labels to each user question. Given the dialogue history, the model infers a label $l_i$ for each $q_i$, covering cues such as emotion, emphasis, speech rate, and tone. Meanwhile, the LLM generates response criteria that specify how the assistant should recognize these cues, track their changes across turns, and respond with appropriate sensitivity.

**Step 3: Style-Controlled Question Re-synthesis.** We convert each paralinguistic label into a natural-language TTS style prompt and synthesize the corresponding user question with Instruct TTS model:

$$q_i^{\text{para}} = \text{TTS}_{\text{inst}}\big(q_i; \text{spk} = \text{ID}_k^{\text{user}}, \text{style} = p_i^{\text{user}}\big), \tag{3}$$

where $p_i^{\text{user}}$ denotes the style prompt derived from the paralinguistic label $l_i$. The same user reference speech is used across turns within a dialogue, ensuring speaker-style consistency while injecting diverse paralinguistic cues into the acoustic signal.

**Step 4: Context-Aware Response Generation.** For each user query, we discard the original response $r_i$ and prompt the LLM with the dialogue history, the paralinguistic label $l_i$, and the generated response criteria. The LLM then produces a new textual response $r_i^{\text{new}}$ that explicitly reacts to both the semantic content and the paralinguistic cues. The response is then synthesized by the Instruct TTS model:

$$r_i^{\text{neu}} = \text{TTS}_{\text{inst}}\big(r_i^{\text{new}}; \text{spk} = \text{ID}_k^{\text{asst}}, \text{style} = p_i^{\text{asst}}\big), \tag{4}$$

where $p_i^{\text{asst}}$ denotes an assistant-side style prompt for natural delivery. The assistant reference speech is also kept fixed across turns to preserve a consistent speaker style.

Each sample of the final output dataset of this stage is a speech pair $(q_i^{\text{para}}, r_i^{\text{neu}})$, where the user query contains explicit paralinguistic information and the assistant response is generated to appropriately reflect that expressive context.

**Dataset Quality Verification** We verify the quality of both synthesized datasets. We compute speaker-embedding cosine similarity across turns within each dialogue to confirm speaker consistency. We also assess speech naturalness with the UTMOS predictor (Saeki et al., 2022), obtaining MOS scores above 4.0.

### 3.2 MODEL ARCHITECTURE

#### 3.2.1 GENERAL SPEECH LLMS INFERENCE

During multi-turn spoken dialogue, when the model receives a new user question, all speech segments in the dialogue context, including both historical speech turns and the current user speech, are first converted into speech representations by the speech encoder. Let $\mathcal{S}_{<t}$ and $\mathcal{T}_{<t}$ denote the historical speech and text context before turn $t$, and let $s_t$ denote the current user speech. The speech encoder produces

$$Z_{<t} = \text{Enc}(\mathcal{S}_{<t}), \qquad z_t = \text{Enc}(s_t). \tag{5}$$

The LLM backbone then takes the interleaved sequence of historical text representations, historical speech representations, and the current speech representation as input:

$$H_t = \text{LLM}\big(\text{Interleave}(\mathcal{T}_{<t}, Z_{<t}, z_t)\big), \tag{6}$$

where $H_t = \{h_{t,1}, \cdots, h_{t,N_t}\}$ denotes LLM hidden states, and $N_t$ is the sequence length at turn $t$.

In end-to-end SLMs, hidden states at different positions are typically routed to different output modules. Some positions are passed to the LLM head to generate text tokens, while speech-related positions are sent to the speech decoder to produce discrete speech tokens, which are then converted into waveform by the speech vocoder. We denote the indices of LLM hidden-state positions routed to the speech decoder as $\mathcal{I}_{\text{speech}}$.

**Why Multi-Turn Spoken Dialogue Is Challenging.** This inference procedure creates a bottleneck in multi-turn spoken dialogue. For each speech position, the information provided to the speech decoder is only the current LLM hidden state $h_{t,j}$. This representation is expected to simultaneously contain the semantic content of the dialogue history, paralinguistic cues from historical speech, and speaker-style information required for consistent responses. Modeling all these factors solely within the LLM backbone imposes a heavy burden, especially when the relevant context is distributed across speech and text modalities.

#### 3.2.2 SPEECHANCHOR

To address this limitation, we introduce the SpeechAnchor module. We adopt **Gated DeltaNet** (Yang et al., 2024) as the SpeechAnchor module, which is a gated state-space architecture that maintains contextual information through a recurrent state-space matrix and supports efficient long-range dependency modeling.

At each speech generation position, SpeechAnchor takes as input the hidden states from historical turns and those already generated in the current turn that will be processed by the speech decoder. Formally, for the $j$-th speech position in turn $t$, we define the speech-modality context as

$$\mathcal{C}_{t,j} = \{h_{\tau,k} \mid \tau < t, k \in \mathcal{I}_{\text{speech}}\} \cup \{h_{t,k} \mid k < j, k \in \mathcal{I}_{\text{speech}}\}. \tag{7}$$

Using the current speech hidden state as the query, SpeechAnchor integrates this context and produces an anchor representation:

$$r_{t,j} = \text{SpeechAnchor}(\mathcal{C}_{t,j}, h_{t,j}). \tag{8}$$

The anchor representation is then injected into the speech decoder by adding it to the output of every Transformer layer:

$$u^{\ell}_{t,j} = \tilde{u}^{\ell}_{t,j} + W r_{t,j}, \qquad \ell = 1, \cdots, L, \tag{9}$$

where $\tilde{u}^{\ell}_{t,j}$ is the original output of the $\ell$-th speech-decoder layer, $u^{\ell}_{t,j}$ is the updated layer output, and $W$ is a shared projection from the LLM hidden-state space to the speech-decoder hidden space.

**Why SpeechAnchor Better Handles Spoken Multi-Turn Dialogue?** In this way, the output signal of SpeechAnchor module contains semantic and paralinguistic information integrated from the entire audio history, as well as speaker information accumulated from all previous turns. The SpeechAnchor module explicitly integrates dialogue history from the speech modality, allowing the speech decoder to receive richer speech-related contextual information. This decoupled design alleviates the burden on the LLM backbone for modeling conversational context entangled across modalities; without it, such information would need to be captured solely by the LLM backbone.

### 3.3 TRAINING STRATEGY

We train SpeechAnchor with a two-stage strategy aligned with our data construction pipeline. In the first stage, the model is trained on `Semantic Consistency Data` to learn multi-turn semantic coherence and stable role-specific speaker identity. In the second stage, we fine-tune the model with `Paralinguistic Consistency Data` and its paired neutral samples from `Semantic Consistency Data`. This paired training encourages the model to distinguish paralinguistic cues from neutral speech and generate responses that are both semantically appropriate and paralinguistically aware.

For both stages, we use the same end-to-end objective. Given input sequence $\mathbf{x}$, target text tokens $\mathbf{y} = \{y_1, \cdots, y_m\}$, and target speech-token groups $\mathbf{s} = \{\mathbf{s}_j\}_{j \in \mathcal{I}_{\text{speech}}}$, the training loss is

$$\mathcal{L} = -\sum_{i=1}^{m} \log P\big(y_i \mid y_{<i}, \mathbf{x}; \theta\big) - \sum_{j \in \mathcal{I}_{\text{speech}}} \log P\big(\mathbf{s}_j \mid \mathbf{s}_{<j}, h_j, r_j; \theta\big), \tag{10}$$

where the first term is the autoregressive text prediction loss, and the second term is the autoregressive speech-token prediction loss. Here, $h_j$ is the LLM hidden state at speech position $j$, $r_j = \mathrm{SpeechAnchor}(\mathcal{C}_j, h_j)$ is the anchor signal injected into the speech decoder, and $\mathbf{s}_j$ denotes the ground-truth speech tokens encoded from the target audio at position $j$.

**Why End-to-End Training Improves Consistency?** With the proposed architecture and consistency-aware data, end-to-end training jointly improves the three forms of consistency. Although the SpeechAnchor module is placed after the LLM in the forward pass, the speech-token loss backpropagates through the speech decoder and SpeechAnchor module into the LLM backbone. Since the target speech tokens are derived from ground-truth audio, this supervision carries both semantic and acoustic information. For ***Semantic*** consistency, this supervision helps the model better understand and integrate multi-turn speech inputs. For ***Paralinguistic*** consistency, paralinguistic-augmented data encourages the model to distinguish acoustic cues from neutral speech and use them for context-aware response generation. For ***Speaker-Style*** consistency, SpeechAnchor processes the same speech-related hidden states used by the speech decoder, allowing the injected anchor representation, together with consistent reference audio within each dialogue, to help maintain stable speaker characteristics across turns.

# 4 EXPERIMENT

## 4.1 EXPERIMENTAL SETUP

**Base Models.** We conduct experiments on two strong end-to-end SLMs: MiMo-Audio-7B (Zhang et al., 2025) and Kimi-Audio-7B (Ding et al., 2025). We select them for their strong performance, open-source reproducibility, and distinct system designs, enabling us to evaluate SpeechAnchor across different end-to-end SLM architectures.

**Training Data.** We use Magpie-Ultra[2] as the source multi-turn textual dialogue dataset for filtering and construction. Magpie-Ultra served as the core multi-turn instruction dataset during SmolLM2 post-training. It was constructed through teacher-model generation under controlled system prompts, followed by rigorous quality filtering and semantic deduplication. Prior work shows that it provides strong supervision for multi-turn instruction following, response quality, reasoning, and mathematics (Allal et al., 2025). We employ `GPT-5.2-2025-12-11`[3] to rewrite responses into speech-friendly expressions, filter questions, annotate paralinguistic labels, generate consistency criteria, and produce context-aware responses. For speech synthesis, we use `CosyVoice3`[4] (Du et al., 2025b) and `Gemini-2.5-Pro-Preview-TTS`[5] as Instruct TTS models.

## 4.2 BENCHMARKS

We conduct experiments on three open-source, multi-turn speech-to-speech benchmarks:

**VoiceAssistant-Eval** (Wang et al., 2025) evaluates multi-turn voice assistant interactions, with dialogues of up to 7 turns, in terms of coherence, relevance, and factual correctness. We use its Speaking-MultiRound subset and report three official metrics: *Content* measures response correctness with an LLM judge after speech transcription, *Consistency* measures agreement between text and spoken outputs, and *Speech* measures generated speech quality with UTMOS.

**MTalk-Bench** (Du et al., 2025a) evaluates multi-turn spoken dialogue across nine common interpersonal communication scenarios and three evaluation tracks: *Semantic* evaluates dialogue history tracking and contextual reasoning, *Paralinguistic* assesses the perception and expression of paralinguistic information, and *Ambient* evaluates the understanding of surrounding acoustic context. MTalk-Bench uses rubric-based LLM judging, and **we adopt Gemini-2.5-Pro for its highest reported agreement with human judgments.**

[2] `https://huggingface.co/datasets/argilla/magpie-ultra-v0.1`

[3] `https://platform.openai.com/docs/models/gpt-5.2`

[4] `https://huggingface.co/FunAudioLLM/Fun-CosyVoice3-0.5B-2512`

[5] `https://blog.google/innovation-and-ai/technology/developers-tools/gemini-2-5-text-to-speech`

Table 1: Evaluation results on MTalk-Bench.

| Model | Semantic | Para. | Ambient | Avg. |
|---|---|---|---|---|
| Step-Audio-Chat | 62.81 | 55.64 | 56.23 | 58.25 |
| Kimi-Audio-7B | 55.33 | 60.62 | 58.63 | 58.17 |
| MiMo-Audio-7B | 64.45 | 56.96 | 56.96 | 59.46 |
| Qwen2.5-Omni-7B | 60.22 | 58.34 | 64.81 | 61.12 |
| Doubao | **70.02** | 60.31 | 60.63 | 63.69 |
| **SpeechAnchor (Kimi)** | 59.21 | 61.73 | 64.74 | 61.89 |
| **SpeechAnchor (MiMo)** | 64.90 | **62.46** | **68.60** | **65.32** |

Table 2: Evaluation results on VoiceAssistant-Eval.

| Model | Content | Cons. | Speech | Final |
|---|---|---|---|---|
| Kimi-Audio-7B | 55.0 | 88.0 | 49.6 | 24.0 |
| MiMo-Audio-7B | 64.8 | 91.6 | 66.5 | 39.5 |
| LLaMA-Omni2-14B | 64.0 | **92.0** | 69.7 | 41.0 |
| **SpeechAnchor (Kimi)** | 55.3 | 88.0 | 53.3 | 25.9 |
| **SpeechAnchor (MiMo)** | **65.2** | 91.6 | **71.4** | **42.7** |

**SpeechRole-Eval** (Jiang et al., 2025) targets multi-turn speech-based role-playing dialogues. It evaluates whether a model can sustain a stable character identity, persona, speaking style, emotional expression, and factual knowledge throughout extended conversations. SpeechRole adopts an LLM-based pairwise evaluation protocol with Gemini-2.5-Pro.

Beyond multi-turn dialogue, we evaluate our model on four general-purpose speech benchmarks, including **Web-Questions** (Berant et al., 2013), **Llama-Questions** (Nachmani et al., 2023), **Alpaca-Eval** (Li et al., 2023; Taori et al., 2023), and **TriviaQA** (Joshi et al., 2017); more details are provided in Appendix A.2.2.

Table 3: Evaluation results on the SpeechRole benchmark. Definitions of all metrics are provided in Appendix A.2.1

| Model | IA | CC | SF | SN | PC | EA | PeC | KC | All |
|---|---|---|---|---|---|---|---|---|---|
| GPT-4o-Audio | 85.8 | 96.1 | **90.2** | **89.6** | 82.8 | 84.1 | 80.0 | 86.8 | 86.9 |
| Kimi-Audio-7B | 69.2 | 75.0 | 57.3 | 55.7 | 61.4 | 66.3 | 66.0 | 71.3 | 65.3 |
| Qwen2.5-Omni-7B | 64.4 | 76.7 | 89.0 | 79.9 | 60.8 | 62.0 | 51.1 | 65.2 | 68.6 |
| MiMo-Audio-7B | 87.5 | 94.9 | 72.5 | 70.4 | 77.7 | 83.9 | 83.5 | 90.2 | 82.6 |
| **SpeechAnchor (Kimi)** | 73.6 | 77.4 | 63.4 | 61.8 | 67.6 | 71.5 | 71.5 | 74.7 | 70.2 |
| **SpeechAnchor (MiMo)** | **93.1** | **97.9** | 80.2 | 78.2 | **85.5** | **90.5** | **90.4** | **94.5** | **88.8** |

## 4.3 Results

### 4.3.1 Multi-turn Spoken Dialogue

We evaluate SpeechAnchor on three multi-turn spoken dialogue benchmarks covering complementary aspects of cross-turn consistency. Across all benchmarks, both SpeechAnchor variants consistently improve over their corresponding backbones, demonstrating robust gains across different models and evaluation settings.

On MTalk-Bench (Du et al., 2025a), SpeechAnchor (MiMo) achieves the best average score and the strongest paralinguistic and ambient performance. On SpeechRole-Eval (Jiang et al., 2025), it achieves the best overall score with substantial gains in identity, emotion, personality, and knowledge consistency. The remaining gaps, including Semantic on MTalk-Bench and Style Fidelity and Speech Naturalness on SpeechRole-Eval, occur against systems with substantially different backbone capacities and model scales.

On VoiceAssistant-Eval (Wang et al., 2025), SpeechAnchor (MiMo) achieves the highest final score, improving both Content and Speech. Its Consistency score remains at 91.6, close to the best score of 92.0 from LLaMA-Omni2-14B (Fang et al., 2025). Since this metric measures agreement between textual and spoken outputs rather than multi-turn response quality, the unchanged score shows that SpeechAnchor improves response and speech quality without compromising text-speech consistency.

### 4.3.2 General SpokenQA

As shown in Table 4, SpeechAnchor achieves comparable or slightly better performance than the backbone models across all four benchmarks. This preservation mainly comes from two design choices. First, the anchor module provides multi-turn context to the speech decoder without altering the core LLM backbone, thus retaining the backbone's semantic reasoning ability. Second, the two-stage training strategy focuses on consistency modeling while preserving the backbone's learned general representations. These results show that SpeechAnchor improves multi-turn consistency

Table 4: Results on general SpokenQA benchmarks.

| Model | Llama-Q | Web-Q | Alpaca | TriviaQA |
|---|---|---|---|---|
| GLM-4-Voice-9B | 50.7 | 15.9 | 4.1 | 23.3 |
| Kimi-Audio-7B | 54.6 | 29.6 | 5.6 | 31.8 |
| LLaMA-Omni-7B | 49.0 | 23.7 | 4.8 | 27.0 |
| LLaMA-Omni2-14B | 62.7 | 37.1 | 6.7 | 38.1 |
| MiMo-Audio-7B | 69.1 | 37.5 | 7.1 | 40.2 |
| **SpeechAnchor (Kimi)** | 55.1 | 29.9 | 5.8 | 32.2 |
| **SpeechAnchor (MiMo)** | **69.7** | **37.9** | **7.3** | **40.7** |

Table 5: Ablation study on MTalk-Bench using MiMo-Audio as the backbone.

| Variant | Semantic | Para. | Ambient | Avg |
|---|---|---|---|---|
| **SpeechAnchor** | **64.90** | **62.46** | **68.60** | **65.32** |
| -Arch | 62.01 | 61.24 | 67.79 | 63.68 |
| -Data | 41.23 | 51.17 | 42.73 | 45.04 |

without sacrificing general spoken interaction ability. To quantify uncertainty, we additionally report 95% confidence intervals from 10,000 bootstrap resamples for all main results in Appendix A.4.

# 5 Analysis

## 5.1 Effect of SpeechAnchor

SpeechAnchor incorporates two key components: the anchor module for multi-turn context modeling and a consistency-aware data synthesis pipeline for training data construction. To assess the effectiveness of each component, we conduct ablation studies by removing them individually. Specifically, we compare three variants: **(1) SpeechAnchor**, the full framework combining both components; **(2) -Arch**, removing the anchor module and training the backbone on our synthesized multi-turn data; and **(3) -Data**, removing our synthesized data and training the anchor module on existing public multi-turn speech dialogue datasets. For the '-Data' variant, we use five representative public multi-turn conversation and emotion-rich dialogue datasets, including *DeepDialogue* (Koudounas et al., 2025), *IEMOCAP* (Busso et al., 2008), *MultiDialog* (Park et al., 2024), *DailyTalk* (Lee et al., 2023), and *MELD* (Poria et al., 2019).

As shown in Table 5, the full SpeechAnchor framework performs best across all MTalk-Bench dimensions, validating both components. Removing the anchor module consistently degrades performance, highlighting the benefit of speech-modality history for cross-turn context modeling. Replacing our synthesized data with public multi-turn speech datasets causes larger drops, showing that the gains stem from consistency-aware data synthesis rather than simply adding more multi-turn speech data. Together, the two components are complementary: the data provides targeted supervision, while the anchor module enables effective use of dialogue history.

Table 6: Analysis of training strategies on SpeechRole using MiMo-Audio as the backbone.

| Variant | IA | CC | SF | SN | PC | EA | PeC | KC | All |
|---|---|---|---|---|---|---|---|---|---|
| Stage 1 | 90.4 | 96.3 | 74.3 | 72.1 | 82.0 | 84.8 | 85.7 | 92.8 | 84.8 |
| Stage 2 | 86.1 | 92.7 | 78.1 | 76.8 | 79.5 | 89.1 | 84.8 | 89.4 | 84.6 |
| Stage 1 $\cup$ Stage 2 | 89.6 | 95.4 | 77.3 | 75.6 | 82.4 | 88.1 | 87.8 | 92.3 | 86.1 |
| Stage 1 $\rightarrow$ Stage 2 | **93.1** | **97.9** | **80.2** | **78.2** | **85.5** | **90.5** | **90.4** | **94.5** | **88.8** |

## 5.2 Effectiveness of Training Strategy

We ablate the two-stage curriculum on SpeechRole to assess its impact on multi-turn consistency. Specifically, we compare four variants: (1) **Stage 1**, training only on `Semantic Consistency Data` for semantic consistency; (2) **Stage 2**, training only on `Paralinguistic Consistency Data` for paralinguistic consistency; (3) **Stage 1** $\cup$ **Stage 2**, merging both datasets into a single training stage; and (4) **Stage 1** $\rightarrow$ **Stage 2**, the full two-stage curriculum used in SpeechAnchor.

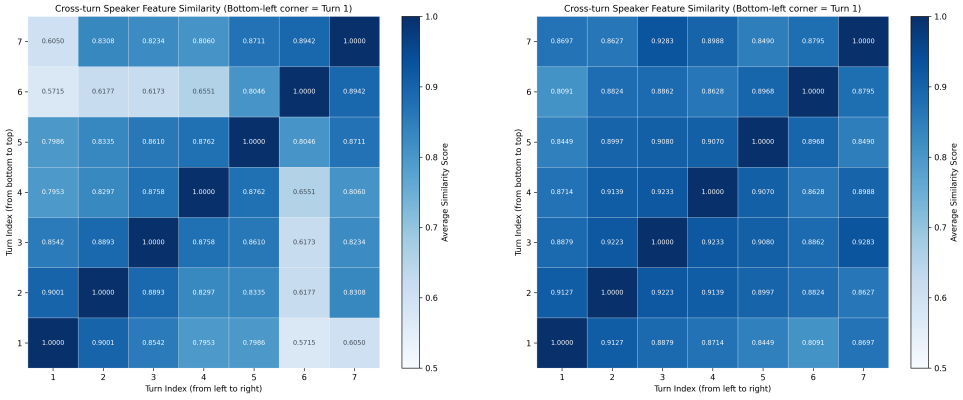
As shown in Table 6, Stage 1 mainly strengthens semantic and speaker consistency, while Stage 2 further improves paralinguistic consistency. Their sequential combination achieves the strongest overall performance, whereas direct data mixing is less effective. This supports our motivation that robust multi-turn spoken dialogue benefits from semantic and speaker grounding before paralinguistic refinement.

### 5.3 Speaker Identity Consistency in Multi-Turn Dialogue

To directly quantify whether SLMs preserve a stable speaking identity as dialogue progresses, we conduct a cross-turn consistency analysis on the three multi-turn benchmarks used in our main experiments. For each evaluated conversation, we extract the assistant's generated speech at every turn and measure pairwise consistency between any two assistant turns within the same conversation, focusing specifically on *speaker identity*.

Given a conversation with $T$ assistant turns, we consider all turn pairs $(i, j)$ with $1 \leq i, j \leq T$. For each pair, we measure speaker-identity consistency using the pretrained speaker verification model **spkrec-ecapa-voxceleb**[6] (Desplanques et al., 2020; Ravanelli et al., 2021). Specifically, we extract speaker embeddings from the assistant speech at turns $i$ and $j$, and compute their cosine similarity. A higher similarity score indicates stronger preservation of speaker identity across turns. We repeat this procedure for all conversations and aggregate the scores for each $(i, j)$ pair across the dataset.

For visualization, we merge conversations from all three benchmarks and plot the aggregated pairwise similarities as a single heatmap. Each entry at row $i$ and column $j$ denotes the average speaker-identity similarity between the assistant's speech at turns $i$ and $j$ across all conversations. As shown in Figure 2, SpeechAnchor consistently achieves higher pairwise speaker-identity similarity than the baseline across the similarity matrix. This result demonstrates that SpeechAnchor better preserves speaker identity throughout multi-turn interactions.




(a) Baseline (b) SpeechAnchor

Figure 2: Speaker consistency of responses across different turns in multi-turn dialogue. The x-axis and y-axis represent different turns, and darker colors indicate higher similarity.

## 6 Conclusion

In this paper, we propose SpeechAnchor to improve multi-turn dialogue consistency in Speech Large Language Models (SLMs). Addressing both architectural and data bottlenecks, SpeechAnchor introduces a lightweight multi-turn anchor module and a consistency-aware data synthesis pipeline. Experiments on seven benchmarks show that SpeechAnchor improves semantic coherence, paralinguistic awareness, and speaker-style preservation in strong SLMs, while maintaining single-turn interaction quality.

## 7 Limitations

This work mainly validates SpeechAnchor on a limited set of representative SLM backbones and synthesized multi-turn speech dialogue data. Although the results demonstrate consistent improvements, further experiments on a broader range of model architectures and larger-scale training data would provide a more comprehensive understanding of the method's generality and scalability. In future work, we plan to scale up both the model coverage and data size, and further investigate how SpeechAnchor performs under more diverse dialogue scenarios and deployment settings.

[6] https://huggingface.co/ravanelli2021speechbrain/spkrec-ecapa-voxceleb

### AI USE STATEMENT

In this work, we used generative AI tools to assist with data construction, including data rewriting, filtering, annotation, response generation, and speech synthesis. Generative AI tools were also used to assist with the implementation and debugging of research code, as well as manuscript drafting and language editing. We did not use generative AI tools to propose the research idea or design the methodology, nor did we use them to analyze experimental results or draw scientific conclusions. AI-assisted outputs were reviewed and verified by the authors. Generated data underwent quality filtering and manual inspection, AI-generated or AI-modified code was executed and tested by the authors, and AI-assisted manuscript text was reviewed by the authors. All reported experimental results, scientific claims, and conclusions were independently verified and approved by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work by providing detailed descriptions of the proposed method, experimental settings, implementation details, and evaluation protocols in the main paper and supplementary materials. The supplementary materials include the source code required to reproduce our experiments, and we will publicly release the code after the review process to further facilitate verification and future research. The content of this paper can be completely reproduced.

## REFERENCES

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1533–1544, 2013.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, et al. Slam-omni: Timbre-controllable voice interaction system with single-stage training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2262–2282, 2025.

Xize Cheng, Dongjie Fu, Xiaoda Yang, Minghui Fang, Ruofan Hu, Jingyu Lu, Bai Jionghao, Zehan Wang, Shengpeng Ji, Rongjie Huang, et al. Omnichat: Enhancing spoken dialogue systems with scalable synthetic data for diverse scenarios. *arXiv preprint arXiv:2501.01384*, 2025.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, et al. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*, 2024.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

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.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In Helen Meng, Bo Xu, and Thomas Fang Zheng (eds.), *Interspeech 2020*, pp. 3830–3834. ISCA, 2020.

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

Yuhao Du, Qianwei Huang, Guo Zhu, Zhanchen Dai, Shunian Chen, Qiming Zhu, Le Pan, Minghao Chen, Yuhao Zhang, Li Zhou, et al. Mtalk-bench: Evaluating speech-to-speech models in multi-turn dialogues via arena-style and rubrics protocols. *arXiv preprint arXiv:2508.18240*, 2025a.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025b.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*, 2024.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. Llama-omni 2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18617–18629, 2025.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

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

Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, et al. Audio multichallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction. *arXiv preprint arXiv:2512.14865*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Changhao Jiang, Jiajun Sun, Yifei Cao, Jiabao Zhuang, Hui Li, Baoyu Fan, Tao Ji, Tao Gui, and Qi Zhang. Speechrole: A large-scale dataset and benchmark for evaluating speech role-playing agents. *arXiv preprint arXiv:2508.02013*, 2025.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Heeseung Kim, Che Hyun Lee, Sangkwon Park, Jiheum Yeom, Nohil Park, Sangwon Yu, and Sungroh Yoon. Does your voice assistant remember? analyzing conversational context recall and utilization in voice interaction models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8984–9014, 2025.

Alkis Koudounas, Moreno La Quatra, and Elena Baralis. Deepdialogue: A multi-turn emotionally-rich spoken dialogue dataset. *arXiv preprint arXiv:2505.19978*, 2025.

Keon Lee, Kyumin Park, and Daeyoung Kim. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`, 2023.

Yu-Xiang Lin, Cheng-Han Chiang, and Hung-yi Lee. Style amnesia: Investigating speaking style degradation and mitigation in multi-turn spoken language models. *arXiv preprint arXiv:2512.23578*, 2025.

Jingwen Liu, Kan Jen Cheng, Jiachen Lian, Akshay Anand, Rishi Jain, Faith Qiao, Robin Netzorg, Huang-Cheng Chou, Tingle Li, Gopala Anumanchipalli, et al. Emo-reasoning: Benchmarking emotional reasoning capabilities in spoken dialogue systems. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2025.

Chengqian Ma, Wei Tao, and Steven Y Guo. C3: A bilingual benchmark for spoken dialogue models exploring challenges in complex conversations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22789–22807, 2025.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm. *arXiv preprint arXiv:2305.15255*, 2023.

Se Park, Chae Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeonghun Yeo, and Yong Ro. Let's go real talk: Spoken dialogue model for face-to-face conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16334–16348, 2024.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

Qwen. Qwen/qwen3-next-80b-a3b-instruct. `https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct`, 2025. Hugging Face model page.

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. PMLR, 2023.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

Qundong Shi, Jie Zhou, Biyuan Lin, Junbo Cui, Guoyang Zeng, Yixuan Zhou, Ziyang Wang, Xin Liu, Zhen Luo, Yudong Wang, et al. Ultraeval-audio: A unified framework for comprehensive evaluation of audio foundation models. *arXiv preprint arXiv:2601.01373*, 2026.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

Ke Wang, Houxing Ren, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Voiceassistant-eval: Benchmarking ai assistants across listening, speaking, and viewing. *arXiv preprint arXiv:2509.22651*, 2025.

Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*, 2024.

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. *arXiv preprint arXiv:2502.17810*, 2025.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. *arXiv preprint arXiv:2412.06464*, 2024.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507, 2021.

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.

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, 2023a.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, et al. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*, 2023b.

# A Additional Experimental Details

## A.1 SpeechAnchor Module Configuration

We instantiate the SpeechAnchor module with Gated DeltaNet (Yang et al., 2024), a gated state-space architecture that maintains contextual information through a recurrent state-space matrix and supports efficient long-range dependency modeling. The output hidden dimension of SpeechAnchor is set to the hidden dimension of the LLM backbone, so that the SpeechAnchor representation can be directly projected and added to the decoder hidden states (Zhang et al., 2025). For the remaining architectural hyperparameters, we follow the `Qwen3-Next-80B-A3B-Instruct` configuration (Qwen, 2025; Team, 2025). Specifically, we adopt its settings for the number of key heads and value heads in the internal multi-head linear aggregation module, the per-head dimensions of the key and value representations, the kernel size of the depthwise one-dimensional convolution used for local temporal mixing, the numerical stability constant in RMSNorm, the learnable DeltaNet discretization bias, and the output projection that maps the aggregated multi-head representation back to the backbone hidden space.

The parameters are initialized as follows: linear transformations are initialized with Xavier initialization (Glorot & Bengio, 2010), convolutional kernels are initialized with Kaiming initialization (He et al., 2015), normalization scale parameters are initialized to 1, the DeltaNet discretization bias $b_\Delta$ is initialized to 1, the log-space transition parameter $A_{\log}$ is sampled from a log-uniform distribution, and the output projection $W_o$ is initialized to 0.

## A.2 Benchmark Details

We provide additional details on the benchmarks used in our main experiments.

### A.2.1 Multi-Turn Speech-to-Speech Dialogue Benchmarks

**VoiceAssistant-Eval.** VoiceAssistant-Eval (Wang et al., 2025) evaluates whether speech language models can provide coherent, relevant, and factually correct responses in voice assistant scenarios. We use its *Speaking-MultiRound* subset, denoted as VoiceAssistant-Eval in our experiments.

VoiceAssistant-Eval reports three official metrics. ***Content*** measures the correctness of the response content by first transcribing the generated speech and then using an LLM judge to evaluate the transcribed answer. ***Consistency*** measures whether the model's textual output matches the transcription of its spoken output, computed with a modified WER after transcription using Whisper-Large-v3[7]. ***Speech*** evaluates the quality of the generated speech using UTMOS, mainly reflecting fluency and naturalness. Following the original benchmark protocol, we also report the final score obtained by combining these three metrics.

**MTalk-Bench.** MTalk-Bench (Du et al., 2025a) evaluates multi-turn spoken dialogue competence under diverse conversational and acoustic conditions. It covers nine high-frequency interpersonal communication scenarios and contains three subsets, namely *Semantic*, *Paralinguistic*, and *Ambient*. These are evaluation tracks rather than different metrics. The ***Semantic*** track focuses on literal semantic understanding, including dialogue history tracking, semantic disambiguation, reasoning, and contextual consistency. The ***Paralinguistic*** track evaluates the perception and expression of non-lexical speech cues, such as emotion, intonation, stress, and speaking rate. The ***Ambient*** track examines the model's ability to handle environmental acoustic context, including background noise, environmental events, and multi-speaker scenarios.

MTalk-Bench adopts a rubric-based absolute scoring protocol, where each sample is evaluated using both general rubrics for shared response qualities, such as relevance and correctness, and specific rubrics targeting the capability of the corresponding track. We use Gemini-2.5-Pro as the LLM judge.

**SpeechRole-Eval.** SpeechRole-Eval (Jiang et al., 2025) targets multi-turn speech-based role-playing dialogues inspired by characters from films, animation, and games. It evaluates whether a

[7] `https://huggingface.co/openai/whisper-large-v3`

model can maintain a stable character identity, persona, speaking style, emotional expression, and background knowledge throughout extended spoken interactions.

SpeechRole-Eval adopts an LLM-based pairwise evaluation protocol with Gemini-2.5-Pro as the judge. It evaluates speech role-playing performance across three dimensions: interaction ability, speech quality and expressiveness, and role-playing fidelity. Specifically, ***Instruction Adherence (IA)*** assesses whether responses follow the task instruction and remain in character; ***Conversational Coherence (CC)*** measures logical consistency with the dialogue context; ***Speech Fluency (SF)*** evaluates fluent delivery with appropriate pacing and minimal disfluencies; ***Speech Naturalness (SN)*** measures how natural and human-like the generated speech sounds; ***Prosodic Consistency (PC)*** assesses whether prosody aligns with the intended speaking style and remains consistent throughout the discourse; ***Emotion Appropriateness (EA)*** evaluates whether emotional expression matches the dialogue context and character state; ***Personality Consistency (PeC)*** measures consistency with the character's personality traits; and ***Knowledge Consistency (KC)*** assesses whether responses remain grounded in the character's established background, knowledge, and relationships. We report all eight metrics together with the overall score.

### A.2.2 Single-Turn SpokenQA Benchmarks

To evaluate whether our method preserves general speech capabilities beyond multi-turn dialogue, we further conduct evaluation on four general-purpose speech benchmarks following the protocol of UltraEval-Audio (Shi et al., 2026). These benchmarks mainly test spoken question answering, factual knowledge, instruction following, and response quality in speech-based settings.

**Web-Questions.** Web-Questions (Berant et al., 2013) is derived from a text-based web question answering dataset and is converted into a spoken QA format. It evaluates the model's ability to understand spoken questions and produce accurate answers to knowledge-intensive queries.

**Llama-Questions.** Llama-Questions (Nachmani et al., 2023) is one of the earliest benchmarks specifically designed for spoken question answering. It tests whether a model can directly process spoken user questions and generate appropriate spoken responses, making it a useful benchmark for evaluating end-to-end spoken QA ability.

**TriviaQA.** TriviaQA (Joshi et al., 2017) is transformed from an open-domain text question answering dataset into a speech-based evaluation setting. It covers a wide range of factual domains and therefore evaluates both spoken comprehension and factual response generation.

**Alpaca-Eval.** Alpaca-Eval (Li et al., 2023) extends the AlpacaEval protocol (Taori et al., 2023) to the spoken instruction-following setting. Unlike the QA-style benchmarks above, it focuses more on response quality, coherence, and user preference. It uses LLM-based pairwise judgments to compare model outputs, providing a complementary evaluation of general spoken instruction-following ability.

## A.3 Training Details

**Training Data** The **Semantic Consistency Data** comprises 80,000 training instances and 10,717.4 hours of speech, while the **Paralinguistic Consistency Data** comprises 20,000 instances and 2,660.7 hours of speech, yielding 100,000 instances and 13,378.1 hours in total.

**Training Config** All experiments were conducted on 8 NVIDIA H100 80GB GPUs. Training is performed in two stages. Stage 1 is trained for one epoch with an effective global batch size of 128, resulting in 625 optimization steps, and uses an initial learning rate of $2 \times 10^{-5}$. Stage 2 is trained for two epochs with an effective global batch size of 16, resulting in 2,500 optimization steps, and uses an initial learning rate of $1 \times 10^{-4}$. Both stages use AdamW (`adamw_torch`) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay of 0.01, and gradient clipping at 1.0. We adopt a cosine learning-rate schedule with a 3% warmup ratio and set the random seed to 42.

## A.4 Bootstrap Confidence Intervals

To quantify uncertainty in the main results, we compute 95% confidence intervals using 10,000 bootstrap resamples over matched evaluation instances. We use complete dialogues as the resampling unit for the three multi-turn benchmarks and individual examples for the four single-turn benchmarks. The resulting confidence intervals are reported in Table 7.

Table 7: Main results with 95% bootstrap confidence intervals. Values are reported as score [95% CI].

| Model | MTalk-Bench | VoiceAssistant-Eval | SpeechRole-Eval | Llama-Q | Web-Q | Alpaca-Eval | TriviaQA |
|---|---|---|---|---|---|---|---|
| Kimi-Audio-7B | 58.17 [56.87, 59.07] | 24.00 [23.34, 24.66] | 65.30 [64.08, 66.52] | 54.60 [53.50, 55.70] | 29.60 [28.72, 30.48] | 5.60 [5.26, 5.94] | 31.80 [30.88, 32.72] |
| **SpeechAnchor (Kimi)** | 61.89 [60.63, 62.95] | 25.90 [25.22, 26.58] | 70.20 [69.02, 71.38] | 55.10 [54.02, 56.18] | 29.90 [29.03, 30.77] | 5.80 [5.45, 6.15] | 32.20 [31.29, 33.11] |
| MiMo-Audio-7B | 59.46 [58.48, 60.34] | 39.50 [38.80, 40.20] | 82.60 [81.50, 83.70] | 69.10 [68.08, 70.12] | 37.50 [36.65, 38.35] | 7.10 [6.74, 7.46] | 40.20 [39.30, 41.10] |
| **SpeechAnchor (MiMo)** | 65.32 [65.04, 65.59] | 42.70 [41.98, 43.42] | 88.80 [87.74, 89.86] | 69.70 [68.69, 70.71] | 37.90 [37.06, 38.74] | 7.30 [6.94, 7.66] | 40.70 [39.81, 41.59] |

Table 8: Ablation study on MTalk-Bench with 95% bootstrap confidence intervals, using MiMo-Audio as the backbone. Values are reported as score [95% CI].

| Variant | Semantic | Para. | Ambient | Avg |
|---|---|---|---|---|
| **SpeechAnchor** | **64.90 [64.54, 65.25]** | **62.46 [62.16, 62.75]** | **68.60 [68.34, 68.86]** | **65.32 [65.04, 65.59]** |
| -Arch | 62.01 [61.65, 62.35] | 61.24 [60.95, 61.52] | 67.79 [67.54, 68.03] | 63.68 [63.41, 63.94] |
| -Data | 41.23 [40.78, 41.67] | 51.17 [50.76, 51.57] | 42.73 [42.34, 43.11] | 45.04 [44.71, 45.37] |

Table 9: Inference overhead of SpeechAnchor using MiMo-Audio-7B as the backbone.

| Metric | MiMo-Audio-7B | SpeechAnchor (MiMo) | Relative Overhead |
|---|---|---|---|
| Per-step Decoding Time | 1.000× | ∼1.012× | ∼+1.2% |
| Full-turn Decoding Time | 1.000× | ∼1.015× | ∼+1.5% |
| GPU Memory Footprint | 1.000× | 1.007× | +0.7% |

Across both backbones, the confidence intervals are consistent with the main findings: SpeechAnchor improves performance on the three multi-turn benchmarks while maintaining comparable performance on the four general SpokenQA benchmarks.

**Ablation Study.** We further apply the same bootstrap procedure to the ablation study on MTalk-Bench. Specifically, we perform 10,000 bootstrap resamples using complete dialogues as the resampling unit and recompute the *Semantic*, *Paralinguistic*, *Ambient*, and average scores for each variant. The resulting 95% confidence intervals are reported in Table 8. The full SpeechAnchor model consistently achieves the highest point estimates across all dimensions, and the bootstrap intervals are consistent with the ablation trends reported in Table 5, further supporting the contributions of both the anchor module and the consistency-aware data synthesis pipeline.

### A.5 Inference Efficiency

We profile MiMo-Audio-7B and SpeechAnchor (MiMo) under the same inference setting to evaluate the runtime cost of explicit speech-history modeling. As shown in Table 9, SpeechAnchor introduces only 1.2% additional per-step decoding time and 1.5% full-turn latency. Its 67.4M additional BF16 parameters require approximately 128.6 MiB, while the recurrent state occupies only 1.06 MiB at batch size 1, increasing the total model memory from 18.846 GiB to 18.972 GiB (+0.7%).

Despite this marginal overhead, SpeechAnchor improves the MTalk-Bench average score by 5.86 points and SpeechRole overall score by 6.2 points over MiMo-Audio-7B. These results demonstrate a favorable performance-cost trade-off, showing that the lightweight recurrent pathway enables explicit speech-history modeling with little additional inference cost.

## B CASE STUDY

Figure 3 and Figure 4 present case studies comparing SpeechAnchor (MiMo) with MiMo-Audio in multi-turn spoken dialogues. As shown in Figure 3, SpeechAnchor demonstrates superior paralinguistic consistency by generating responses that appropriately align with the speaker's emotional tone and acoustic characteristics across turns, whereas MiMo-Audio produces contextually misaligned paralinguistic outputs. Figure 4 further illustrates that SpeechAnchor maintains better semantic consistency throughout the conversation, providing coherent and contextually grounded responses that effectively leverage prior dialogue history. These cases highlight the effectiveness of our anchor mechanism in capturing and utilizing cross-turn information for both paralinguistic and semantic dimensions.

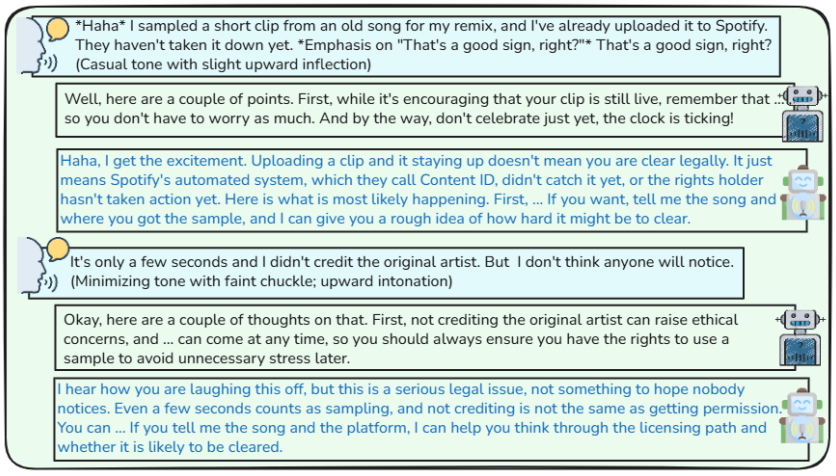



Figure 3: Case study on paralinguistic consistency in multi-turn spoken dialogue: SpeechAnchor responses (blue) contrasted with MiMo-Audio responses (black).

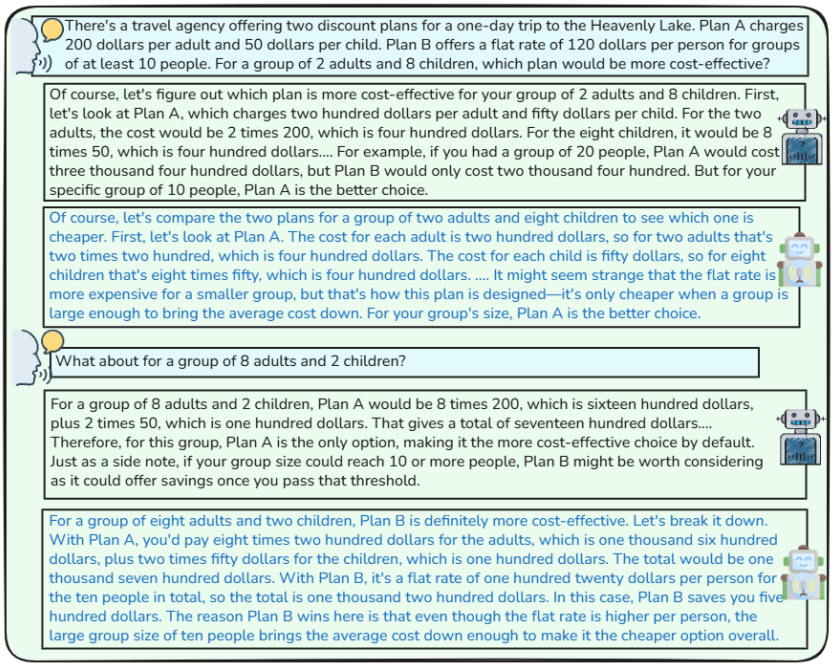



Figure 4: Case study on semantic consistency in multi-turn spoken dialogue: SpeechAnchor responses (blue) contrasted with MiMo-Audio responses (black).

## C PROMPTS

We present the prompts used in our data synthesis pipeline. These prompts guide the LLM-based stages for speech-oriented response rewriting and paralinguistic consistency data construction.

**Prompt for Speech-Friendly Response Rewriting.**

### Task
I am rewriting assistant responses in multi-turn dialogues so that they are more suitable for TTS speech synthesis. You will be given a multi-turn conversation. Your task is to rewrite ONLY the assistant responses. Do NOT rewrite, delete, summarize, or modify the user questions.

### Rewrite Goal
Convert the assistant responses into natural, speech-friendly expressions while preserving the original meaning, intent, tone, and factual content.

### Rewrite Rules
- Remove all Markdown formatting symbols, such as *, **, #, -, >, ‘, and table formatting.
- Replace speech-unfriendly punctuation with natural spoken punctuation. For example, convert em dashes, semicolons, excessive ellipses, and complex punctuation into commas or periods where appropriate.
- Expand or clarify abbreviations when needed for speech. For example, convert “e.g.” to “for example”, “i.e.” to “that is”, and “ICU” to “I-C-U”.
- Keep the response fluent, conversational, and easy to read aloud.

### Conversation to Rewrite
<User Question in turn1>: **{turn1_user}** <Assistant Response in turn1>: **{turn1_assistant}**
<User Question in turn2>: **{turn2_user}** <Assistant Response in turn2>: **{turn2_assistant}**
<User Question in turn3>: **{turn3_user}** <Assistant Response in turn3>: **{turn3_assistant}**

### Output Format
Return only the rewritten assistant responses in the following format:

<Rewritten Assistant Response in turn1>: ...
<Rewritten Assistant Response in turn2>: ...
<Rewritten Assistant Response in turn3>: ...

**Prompt for Paralinguistic-Aware Dialogue Filtering**

### Reference Case 1
<Question in turn1>: ...
<Question in turn2>: ...
<Desired Criteria for Model Responses>: ...

### Reference Case 2
<Question in turn1>: ...
<Question in turn2>: ...
<Desired Criteria for Model Responses>: ...

### Reference Case 3
<Question in turn1>: ...
<Question in turn2>: ...
<Desired Criteria for Model Responses>: ...

### Conversation that needs to be judged
<User Question in turn1>: **{turn1_user}**
<User Question in turn2>: **{turn2_user}**
<User Question in turn3>: **{turn3_user}**

### Requests
I am looking to modify some multi-turn conversations, for example, adding paralinguistic information, but I need to determine if they have the potential for such modification.
Please strictly classify "Conversation that needs to be judged" as YES or NO according to the 3 Reference Cases and this criterion: After adding or altering the paralinguistic information, such as speaker's emotion, stress, intonation, speech rate, and pitch, of questions in a dialogue or across different dialogue rounds, the model's responses will vary significantly depending on the paralinguistic information provided in each turn.
If "Conversation that needs to be judged" fails to meet the criterion above, judge as NO; otherwise, judge as YES. Please output only the final decision in the exact format `\boxed{YES}` or `\boxed{NO}`, with no additional explanation.

**Prompt for Paralinguistic Cue Annotation.**

### Reference Case 1
<Question in turn1>: ...
<Question in turn2>: ...
<Expected Specific Criteria for Model Responses>: ...

### Reference Case 2
<Question in turn1>: ...
<Question in turn2>: ...
<Expected Specific Criteria for Model Responses>: ...

### Reference Case 3
<Question in turn1>: ...
<Question in turn2>: ...
<Expected Specific Criteria for Model Responses>: ...

### Conversation that needs to be Rewriting
<User Question in turn1>: **{turn1_user}**
<User Question in turn2>: **{turn2_user}**
<User Question in turn3>: **{turn3_user}**

### Requests
I am rewriting some text-based multi-turn conversations. My goal is to add paralinguistic information labels to the "Conversation that needs to be Rewriting" and generate corresponding <Expected Specific Criteria for Model Responses>, creating valuable data for spoken multi-turn dialogue.
Please perform the following tasks:
1. Add paralinguistic information to the conversation, adhering to these requirements:
- Keep the paralinguistic information concise and brief, using only a few short phrases.
- After adding or altering the paralinguistic information, such as speaker's emotion, stress, intonation, speech rate, and pitch, of questions in the dialogue or across different dialogue rounds, the model's responses will vary significantly depending on the paralinguistic information provided in each turn.

- Ensure the paralinguistic information aligns naturally with the textual content. Avoid significant conflicts or contradictions between the words and the paralinguistic cues.
- After adding paralinguistic information to the original questions, ensure the overall conversation contains interpretable underlying meanings or subtext as much as possible.

2. Generate <Expected Specific Criteria for Model Responses> by referencing and emulating the "Reference Cases" after adding the paralinguistic information, adhering to these requirements:
- Provide at least 4 points.
- Based on the specific scenario, describe which paralinguistic features the model should identify in the current turn and what changes have occurred compared to previous turns.
- Based on the specific scenario, describe specifically how the model should adjust its response based on these identified paralinguistic features of the current turn as well as the changes compared to previous turns.
- Based on the specific scenario, describe specifically how the model's response should interpret potential underlying meanings or subtext hidden within the user's paralinguistic cues to demonstrate high emotional intelligence.

### Answer Format
Please brainstorm fully, but strictly follow this format for the final output:
<Paralinguistic Information Labels for turn1_user>: ...
<Paralinguistic Information Labels for turn2_user>: ...
<Paralinguistic Information Labels for turn3_user>: ...
<Expected Specific Criteria for Model Responses>: ...

**Prompt for Paralinguistic-Aware Response Generation.**

### Conversation
<User Question in turn1>: **{turn1_user}**
<User Question in turn2>: **{turn2_user}**
<User Question in turn3>: **{turn3_user}**

### Paralinguistic Information Labels
<Paralinguistic Information Labels for turn1_user>: **{turn1_para_label}**
<Paralinguistic Information Labels for turn2_user>: **{turn2_para_label}**
<Paralinguistic Information Labels for turn3_user>: **{turn3_para_label}**

### Expected Specific Criteria for Model Responses
**{specific_criteria}**

### Requests
Please generate a high-quality response for each dialogue turn by strictly adhering to the provided "Expected Specific Criteria for Model Responses". Your responses must effectively integrate the user's textual content with the corresponding "Paralinguistic Information Labels", as explicitly demanded by the criteria.

### Answer Format
<Model Response to the turn1 user's question>: ...
<Model Response to the turn2 user's question>: ...
<Model Response to the turn3 user's question>: ...