

AZEROS: Extending LLM to Speech with Self-Generated Instruction-Free Tuning

Yiwen Shao^{*†}, Wei Liu^{*}, Jiahong Li^{*}, Tianzi Wang, Kun Wei, Meng Yu, Dong Yu

Tencent AI Lab

yiwenyshao@global.tencent.com

^{*}Equal Contribution, [†]Project Lead

Abstract: Extending large language models (LLMs) to the speech domain has recently gained significant attention. A typical approach connects a pretrained LLM with an audio encoder through a projection module and trains the resulting model on large-scale, task-specific instruction-tuning datasets. However, curating such instruction-tuning data for specific requirements is time-consuming, and models trained in this manner often generalize poorly to unseen tasks. In this work, we first formulate that the strongest generalization of a speech-LLM is achieved when it is trained with **Self-Generated Instruction-Free Tuning (SIFT)**, in which supervision signals are generated by a frozen LLM using textual representations of speech as input. Our proposed SIFT paradigm eliminates the need for collecting task-specific question-answer pairs and yields the theoretically best generalization to unseen tasks. Building upon this paradigm, we introduce **AZEROS** (Auden **Z**ero-instruction-tuned Speech-LLM), which is trained on speech-text pairs derived from publicly available corpora, including approximately 25k hours of speech with ASR transcripts and 3k hours of speech with paralinguistic labels. Built upon Qwen2.5-7B-Instruct, the model updates only two lightweight projection modules (23.8M parameters each), while keeping both the LLM and audio encoders frozen. Despite the minimal training cost and modest data scale, AZEROS achieves state-of-the-art performance on both semantic and paralinguistic benchmarks, including VoiceBench, AIR-Bench Foundation (Speech), and AIR-Bench Chat (Speech).

Code: <https://github.com/AudenAI/Auden/tree/main/examples/azeros>

Model: <https://huggingface.co/AudenAI/azeros/tree/main>

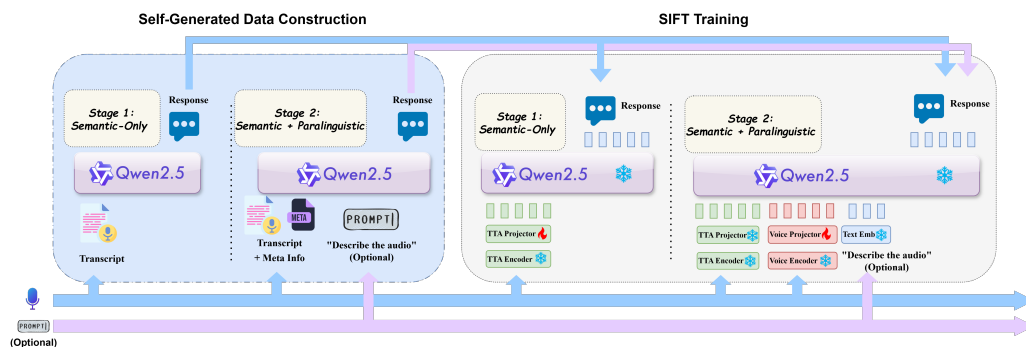


Figure 1: Overview of the proposed AZEROS framework built upon a frozen Qwen2.5-7B-Instruct backbone. The system couples a **Self-Generated Data Construction** pipeline (left) with a two-stage **Self-Generated Instruction-Free Tuning (SIFT)** training procedure (right). In **Stage 1 (Semantic)**, the LLM generates targets from speech transcripts to train a projector on the semantic TTA Liu et al. (2025) encoder. In **Stage 2 (Semantic + Paralinguistic)**, we augment inputs with metadata (e.g., emotion, gender, age); the resulting rich responses train a second projector on a paralinguistic Auden-Voice Huo et al. (2025) encoder. Together, this progressive SIFT paradigm extends the model’s capability from purely semantic to joint semantic-paralinguistic understanding without relying on task-specific instructions.

1 Introduction

Large language models (LLMs) [Mann et al. \(2020\)](#); [Du et al. \(2022\)](#); [Hoffmann et al. \(2022\)](#); [Chowdhery et al. \(2023\)](#); [Chiang et al. \(2023\)](#); [Chung et al. \(2024\)](#); [Achiam et al. \(2023\)](#); [Touvron et al. \(2023\)](#); [Bai et al. \(2023\)](#); [Yang et al. \(2024a\)](#); [Zhang et al. \(2022b\)](#); [Henighan et al. \(2020\)](#); [Kojima et al. \(2022\)](#); [Rae et al. \(2021\)](#); [Jiang et al. \(2023\)](#) have revolutionized the field of Natural Language Processing (NLP). Trained on internet-scale text data through next-token prediction, LLMs acquire vast world knowledge and exhibit powerful language understanding and reasoning capabilities. In recent years, extending LLMs to incorporate additional modalities has become a prominent research trend. By inheriting LLMs’ language generalization abilities, it is expected to develop unified multimodal LLMs (MLLM) [Radford et al. \(2021\)](#); [Liu et al. \(2023\)](#); [Comanici et al. \(2025\)](#); [Yin et al. \(2024\)](#) that can handle multiple tasks across various modalities using a single model.

Among all modalities, speech plays a fundamental role in human communication. A straightforward way to integrate it with LLMs is through a cascaded system. The cascaded speech–LLM pipeline, which first applies automatic speech recognition (ASR) before text-based processing, suffers from error propagation and the loss of non-semantic acoustic information such as prosody and emotion. These limitations motivate the development of end-to-end architectures that process speech directly.

End-to-end speech–LLMs seamlessly connect a speech encoder with an LLM without intermediate text. Forwarding with speech representation mitigates error propagation and could potentially preserve those fine-grained, non-semantic cues. Existing approaches can be broadly categorized into three types [Arora et al. \(2025\)](#): (1) **Pure-speech models**, which handle only speech input and output (e.g., GSLM [Lakhotia et al. \(2021\)](#), AudioLM [Borsos et al. \(2023\)](#)); (2) **Speech-aware models**, which accept both speech and text as input but produce only text output (e.g., Qwen-Audio [Chu et al. \(2023\)](#), SALMONN [Tang et al. \(2023\)](#), Qwen2-Audio [Chu et al. \(2024\)](#), WavLLM [Hu et al. \(2024\)](#)); and (3) **General speech–text models**, which support both modalities at the input and output (e.g., SpeechGPT [Zhang et al. \(2023\)](#), Moshi [Défossez et al. \(2024\)](#) and MiMo-audio [Cici et al. \(2025\)](#)). Recent omni models such as Qwen2.5-Omni [Xu et al. \(2025a\)](#) and Qwen3-Omni [Xu et al. \(2025b\)](#) adopt a thinker–talker architecture, which can be viewed as extending a speech-aware LLM to a general speech–text system by integrating a streaming talker (TTS) module. Motivated by this design, our work focuses on the second category—the speech-aware type—which can be seamlessly upgraded to a general speech–by attaching a streaming TTS system such as CosyVoice [Du et al. \(2024; 2025\)](#).

While recent advances have pushed speech–LLMs toward broader capabilities, their training often relies heavily on large-scale, manually curated instruction data. **We argue that extensive task-specific instruction tuning may be unnecessary if the generalization capacity of a frozen LLM can be fully leveraged.** In this paper, we take a step further by introducing both a theoretical and practical framework for aligning the speech modality with LLMs through self-generated data. Specifically, without relying on any instruction-tuning data, we extend a pretrained LLM into a speech–LLM that preserves its original textual understanding ability while achieving robust comprehension of both semantic and paralinguistic aspects of speech.

Our main contributions are summarized as follows:

- We first formulate **Self-Generated Instruction-Free Tuning (SIFT)** as a training paradigm for speech–LLM models, in which supervision signals are generated by a frozen LLM using **oracle textual representations of speech** as input, without relying on any explicit instructions.
- Building upon the SIFT paradigm, we develop **AZEROS**, a lightweight speech–LLM trained by updating only projection modules while keeping both the LLM and audio encoders frozen.
- Despite minimal training cost and modest data scale, AZEROS achieves state-of-the-art performance on both semantic and paralinguistic benchmarks, including *VoiceBench* [Chen et al. \(2024\)](#) and *AIR-Bench (Speech)* [Yang et al. \(2024b\)](#).

2 Related Work

Training speech-LLMs to achieve general speech understanding remains challenging, as it typically requires large-scale **Task-Specific Instruction Tuning (TSIT)** data. Common speech-domain instruction tasks include automatic speech recognition (ASR), speech translation (ST), speech emotion recognition (SER), and speech question answering (SQA). Qwen-Audio [Chu et al. \(2023\)](#) adopts a multi-task framework that integrates diverse tasks via hierarchical tags, including ASR, ST, SER, language identification (LID), speaker verification (SV), and speaker attribute prediction, etc. Qwen2-Audio [Chu et al. \(2024\)](#) improves this paradigm by using natural-language prompts to unify task instructions. SALMONN [Tang et al. \(2023\)](#) and WavLLM [Hu et al. \(2024\)](#) employ dual encoders and extensive task-specific instruction tuning. It is found that models trained in this way often become biased toward ASR-style responses probably due to the dominance of ASR data in TSIT distribution. To mitigate this, SALMONN applies activation tuning for long-form generative tasks such as storytelling, while WavLLM introduces progressive multi-task mixing to balance single-task overfitting. Nevertheless, each task’s capability still depends on its corresponding instruction dataset, limiting these models to a fixed set of pre-defined tasks rather than achieving a kind of general open-ended speech understanding.

Given the limitations of TSIT, recent work explores leveraging the inherent generalization ability of LLMs to build instruction-following speech models without task-specific supervision. AudioChat-LLaMA [Fathullah et al. \(2024\)](#) aligns speech representations with their transcriptions by ensuring the frozen LLM produces identical responses when fed with either speech or text, thereby enforcing modality consistency. BLSP [Wang et al. \(2023\)](#) achieves similar alignment through a continuation-writing objective, allowing speech and text inputs to elicit equivalent LLM behavior. Although these methods enable zero-shot generalization across many speech tasks, they primarily focus on aligning semantic information and overlook paralinguistic aspects such as emotion or prosody. A follow up work of AudioChat-LLaMA, SpeechEmotionLlama [Kang et al. \(2024\)](#) performs emotion-conditioned response alignment. The resulting system is able to generate more empathetic responses to expressive speech input compared to the baseline [Fathullah et al. \(2024\)](#). Desta2 [Lu et al. \(2024\)](#) explores paralinguistic aspects beyond emotion, establishing that a unified description-based training objective exhibits significant generalization capability. The method utilizes a frozen LLM to generate descriptive targets by integrating both speech transcripts and paralinguistic metadata. Zero-shot instruction following ability on both linguistic and paralinguistic aspects is observed, though not using TSIT data. Desta2.5 [Lu et al. \(2025\)](#) further emphasizes the value of LLM’s self-generated data and extends the model’s capability beyond speech. To fully utilize the available information in textural description, Desta2.5 crafts thousands of prompts instead of a single descriptive prompt used in Desta2. Despite the advantages, Desta series still relies on external ASR models for input transcriptions and cannot natively support speech-based interaction.

3 Methodology

3.1 Motivation: Unlocking the Universal Decoder

Our goal is to adapt a pre-trained Large Language Model (LLM) for speech understanding. We posit that modern LLMs are already “universal decoders”—they possess the inherent ability to follow virtually any instruction if provided with the correct textual information. Therefore, the challenge of multimodal adaptation is not to teach the LLM new tasks, but simply to **align the audio modality with the text modality** so that the LLM can perceive speech as clearly as it perceives text.

However, the dominant paradigm—Instruction Tuning—often undermines this goal. This method trains the entire Speech-LLM on triplets denoted as (x, i, y) , consisting of an audio input x , a specific task instruction i , and a target response y . By optimizing the model to satisfy specific instructions i , we introduce two fundamental limitations that hurt generalization:

1. **Instruction-Induced Feature Collapse:** When the system is optimized for a specific instruction i (e.g., “Transcribe”), the audio encoding layers learn to be “lazy.” They extract only the minimal subset of features from x required for that task (e.g., phonemes) and discard the rest (e.g., emotion, speaker identity) as noise. This effectively creates a **feature filter**, permanently deleting information that might be needed for future, unseen tasks.
2. **Instruction Overfitting (Modality Neglect):** Strong instruction conditioning encourages the model to “cheat.” The MLLM learns spurious correlations between the instruction i and the response y (i.e., modeling $P(y|i)$), effectively bypassing the audio input x . The model becomes “dishonest” to the sensory input, learning to follow templates rather than grounding its generation in the actual audio content.

To visualize this, consider the “**Packed Suitcase**” analogy:

Imagine the MLLM as a traveler packing for a trip (test time). Instruction tuning is like telling them: “You are going to the beach.”

First, they only pack swimsuits (Feature Collapse). Second, because they are focused on the “beach” order, they ignore the weather report (Instruction Overfitting). If they arrive and it is snowing, they fail twice: they lack the right gear, and they didn’t even look at the reality before leaving.

In contrast, an ideal system should “pack everything” and force the traveler to look at the weather. By removing the instruction during training, the model is forced to rely entirely on the audio content to decide what is relevant.

We argue that to leverage the LLM’s true generalizability, we must abandon task-specific training and instead focus on **Instruction-Free Input Alignment**—forcing the system to preserve all information in x so the LLM can handle any instruction i at test time.

3.2 Problem Setup

We formally consider a Speech-LLM framework that maps an audio input $x \in \mathcal{X}$ to a textual output $y \in \mathcal{Y}$. The system relies on a frozen LLM, denoted as g_ϕ , which operates on a continuous embedding space.

The mapping from raw audio to the LLM’s input space is defined as $h_\theta(x)$. While h_θ generally represents the full audio encoding pipeline, in this work we adopt a standard modular architecture where h_θ is the composition of a **pre-trained audio encoder** (e.g., Whisper, HuBERT) and a **learnable projector** (e.g., MLP or Q-Former). For simplicity, **we freeze the audio encoder and optimize only the projector parameters θ** .

Our objective is to learn a mapping h_θ that translates acoustic features into the LLM’s native embedding space such that the audio becomes “readable” by the LLM without altering the LLM itself.

3.3 Self-Generated Data Construction

To achieve this alignment without instruction bias, we need an alignment target that represents the “ideal” interpretation of the audio. We conceptualize this as a semantic equivalent of the audio’s text description.

Let $\tilde{h}(x)$ denote the “oracle” textual representation of an audio input x (e.g., the embedding of its ground-truth transcript). To create a supervision signal, we utilize the frozen LLM g_ϕ to generate target responses under arbitrary instruction i :

$$y = g_\phi(\tilde{h}(x), i). \quad (1)$$

For this process to define a valid alignment target, we pursue two key **Design Goals**:

1. **Goal 1: Maximizing Information in $\tilde{h}(x)$.** We aim for the oracle text representation $\tilde{h}(x)$ to serve as a comprehensive substitute for the raw audio x . Ideally, it captures diverse details—such as content, emotion, and speaker style—to ensure the text embedding represents the audio signal as fully as possible.
2. **Goal 2: Self-Elicit Capability.** The frozen LLM should **intrinsically** generate a response y that covers the full information within $\tilde{h}(x)$ without needing an instruction ($i = \emptyset$).

Based on whether Goal 2 is satisfied, our paradigm branches into two distinct strategies:

- **Method A: SIFT (Optimal).** If the LLM successfully self-elicits the content (Goal 2 Satisfied), we proceed with **Self-Generated Instruction-Free Tuning (SIFT)**. This is the optimal setting where $i = \emptyset$, forcing pure input alignment.
- **Method B: SIT (Fallback).** If the LLM requires prompting to output the full information (Goal 2 Not Satisfied), we adopt **Self-Generated Instruction Tuning (SIT)**. For example, we can use a generic prompt “Describe the audio” to elicit the target y .

Under this unified perspective, prior works using similar *self-generated* ideas can be viewed as constrained special cases of our paradigm, as detailed in Table 1.

Table 1: Unified Comparison of Alignment Strategies. We frame prior works as constrained special cases within our paradigm: AudioChatLlama Fathullah et al. (2024) is limited by information sufficiency (transcript only), while DESTA-2/2.5 Lu et al. (2024; 2025) relies on rigid instructions and auxiliary ASR inputs. SIFT overcomes these bottlenecks by satisfying both Information Maximization and Self-Elicit goals, achieving pure instruction-free alignment.

Method	Oracle Info $\tilde{h}(x)$ (Goal 1)	Instruction I (Goal 2)	Limitations/Characteristics
AudioChatLlama	Transcript Only (e.g., <i>The weather is great!</i>)	Null (\emptyset)	Semantic only alignment, ignores other info.
DESTA-2/2.5	Transcript + Attributes Tags (e.g., <i>The weather is great! (Gender: Female, Emotion: Happy...)</i>)	Required (e.g., “What can you hear from the audio?”)	Still requires instruction I ; Partial alignment , requires ASR text appended at inference.
SIT (Fallback)	Full Info , either in attributes format (e.g., <i>The weather is great! (Gender: Female, Emotion: Happy...)</i>) or natural description (e.g., <i>A woman says happily: “The weather is great!” ...</i>)	Required (e.g., <i>Describe all the information you can hear.</i>)	Elicit all info in rich response y ; Still relies on instruction I , may introduce instruction bias.
SIFT (Optimal)		Null (\emptyset)	Ideal optimal case No instruction bias. Generalized to any task.

3.4 SIFT: Self-Generated Instruction-Free Tuning

We primarily focus on the optimal setting: SIFT. Our objective is **Input-Level Audio-Text Alignment**: we aim to train the encoder h_θ such that the projected audio features become indistinguishable from the oracle text embeddings from the perspective of the LLM:

$$h_\theta(x) \approx \tilde{h}(x). \quad (2)$$

We achieve this by minimizing the prediction error on the self-generated target y **without** providing any instruction ($i = \emptyset$). The optimization problem is:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathcal{L}_{\text{CE}}\left(y, g_\phi(h_\theta(x), \emptyset)\right). \quad (3)$$

By removing the instruction variable, we force the gradient to flow entirely through the encoder, driving it to bridge the modality gap.

3.5 Why Instruction-Free Generalizes Best

The advantage of SIFT is best understood through the lens of alignment completeness.

1. Instruction Tuning yields Partial Alignment: When training with a specific instruction I (maximizing $P(Y | X, I)$), the model only needs to align the subset of features in X relevant to I (e.g., phonemes for ASR), often ignoring paralinguistic details. This results in a **task-dependent alignment** that fails when the LLM is asked to perform a new task requiring different features.

2. SIFT yields General Input Alignment: By removing the instruction (maximizing $P(Y | X)$), we remove the filter. Because the target y reflects the full content of the oracle text $\tilde{h}(x)$, the only way to minimize the loss is for the encoder to achieve **general alignment**. The projected audio must capture all auditory details to reproduce the unguided textual output.

Mathematically, blocking the instruction channel forces the encoder to solve the hardest alignment problem:

$$h_\theta(x) \xrightarrow{\text{align}} \tilde{h}(x) \quad (\text{Globally}). \quad (4)$$

This produces a “universal” encoder that maps audio into the LLM’s space faithfully. At test time, we can simply plug in any instruction I_{new} , and the frozen LLM—being a universal decoder—will execute it successfully because the input is perfectly aligned.

4 AZeroS: System Overview

To verify the effectiveness of the proposed SIFT paradigm, as illustrated in Figure 1, we build a speech-LLM named AZEROS (Auden **Z**ero-instruction-tuned **S**peech-LLM), implemented within the open-source Auden toolkit¹.

4.1 Model Architecture

AZEROS adopts a typical encoder–projector–LLM architecture, following the design of most multi-modal LLMs.

Large Language Model. We select Qwen2.5-7B-Instruct Yang et al. (2024a) as our backbone LLM. Qwen2.5 is a high-performance open-source language model series trained on large-scale data, and its 7B Instruct variant provides strong instruction-following capability and consistently competitive performance across diverse benchmarks.

Semantic Encoder. We adopt the TTA encoder Liu et al. (2025), a 153M Zipformer-based Yao et al. (2023) model trained on 358k hours of multilingual ASR and translation data, to obtain high-quality semantic representations. The encoder operates on 80-dimensional log-Mel filterbank features computed with a 25 ms window and a 10 ms stride, and produces 768-dimensional feature sequences at a frame rate of 25 Hz after a $4\times$ temporal downsampling.

Paralinguistic Encoder. To capture speaker-specific and paralinguistic cues (e.g., emotion, prosody, speaker identity), we adopt the Auden-Voice encoder Huo et al. (2025), a 153M Zipformer-based model trained on 8k hours of multi-task data covering speaker identity, age, gender, and emotion. The encoder produces 768-dimensional paralinguistic embeddings at a frame rate of 25 Hz.

Projector. Each speech encoder is paired with a linear 23.8M projector that downsamples and maps the encoder features into the LLM embedding space. The projector groups every 4 frames and applies a two-layer MLP with a ReLU activation. This reduces the frame rate to 6.25 Hz, aligning the audio representations with the temporal granularity of text tokens.

4.2 Training Paradigm

Following the self-generated paradigm introduced in Section 3.4, AZEROS integrates data generation and model training into a paired process.

¹<https://github.com/AudenAI/Auden>

Data Generation. Given a speech-text pair $(x, \tilde{h}(x))$, where x denotes the speech signal and $\tilde{h}(x)$ is the textual representation we aim to map (e.g., its original supervision such as transcript or metadata), we use $\tilde{h}(x)$ as the sole input to the LLM to generate a supervision response y . This produces a training sample of the form (x, y) .

Model Training. During training, both the LLM and the audio encoders remain frozen. Only the lightweight projector (23.8M parameters) is updated, allowing it to focus solely on learning the desired mapping from audio representations into the LLM embedding space.

Multi-stage Alignment. To better align both the semantic and paralinguistic information in speech with the LLM, we divide training into two stages. In the first stage, only semantic information is considered: we use the transcript as $\tilde{h}(x)$ and align the TTA encoder alone. In the second stage, $\tilde{h}(x)$ is extended to include both the transcript and the associated meta information, enabling the model to learn paralinguistic mappings as well. Since the semantic component has already been aligned in Stage 1, we freeze all modules except the paralinguistic projector and update only this component during Stage 2.

5 Experimental Setup and Results

5.1 Data Source

Table 2 summarizes all datasets used in our experiments. All datasets are publicly available. Based on the available supervision types, we categorize the data into two sets:

Data-SP. This set comprises 3,548 hours of speech with both semantic transcripts and paralinguistic annotations. For corpora that contain paralinguistic annotations but lack transcripts, we generate transcripts using the TTA ASR model Liu et al. (2025).

Data-S. This set consists of two standard Chinese and English ASR corpora together with the multilingual Common Voice corpus. It contains 22,383 hours of speech in total and provides semantic supervision in the form of ASR transcripts only.

Table 2: Summary of data source used for semantic and paralinguistic alignment in AZeroS.

Dataset	Label	Hours (h)
Data-SP		
IEMOCAP Busso et al. (2008)	gender, emotion, transcript ¹	9
CREMA-D Cao et al. (2014)	gender, age, emotion, transcript ¹	5
MELD Poria et al. (2019)	gender, emotion, transcript ¹	8
RAVDESS Livingstone & Russo (2018)	gender, emotion, transcript ¹	1
TESS Pichora-Fuller & Dupuis (2020)	gender, age, emotion, transcript ¹	1
DailyTalk Lee et al. (2023)	emotion, transcript	21
CommonVoice-en Ardila et al. (2019)	gender, age, transcript	1199
AISHELL-1 Bu et al. (2017)	gender, transcript	150
EmotionTalk Sun et al. (2025)	gender, emotion, transcript	23
CS-Dialogue Zhou et al. (2025)	gender, age, transcript	104
VoxCeleb2 Chung et al. (2018)	gender, transcript ¹	2026
Total		3548²
Data-S		
WenetSpeech Zhang et al. (2022a)	transcript	9685
GigaSpeech Chen et al. (2021)	transcript	9396
CommonVoice Ardila et al. (2019)	transcript	3302
Total		22383

¹ Transcripts are generated by the TTA Liu et al. (2025) model.

² Datasets are rebalanced during training.

5.2 Evaluation Benchmarks

Voicebench Chen et al. (2024). We use VoiceBench to evaluate the semantic understanding ability of AZeroS. VoiceBench is specifically designed to assess speech-LLMs in realistic spoken-instruction scenarios, covering diverse tasks such as general knowledge, reasoning, and safety across both real and synthetic speech. Unlike single-task evaluations such as ASR or speech translation, its diversity enables a more comprehensive assessment of a model’s ability to generalize to **unseen spoken tasks** and instructions within semantic understanding. We follow the official evaluation protocol² and report instruction-following and comprehension accuracy.

AIR-Bench-Speech Yang et al. (2024b). To evaluate paralinguistic understanding and complement the semantic evaluation, we adopt the speech-specific subset of AIR-Bench. AIR-Bench encompasses both semantic and non-semantic aspects of speech; our selection focuses on paralinguistic cues such as prosody, emotion, and speaker characteristics, while still retaining semantic components in speech-

²gpt-4o-2025-01-01-preview is used for evaluation.

only tasks. The *Foundation* part contains multiple-choice questions³, and the *Chat*⁴ part consists of open-ended question-answer pairs, both formatted with text-based instructions appended after the audio input. This setup provides a compact yet comprehensive evaluation of how well the model integrates non-linguistic cues with semantic information in spoken input.

5.3 Stage 1: Semantic Alignment

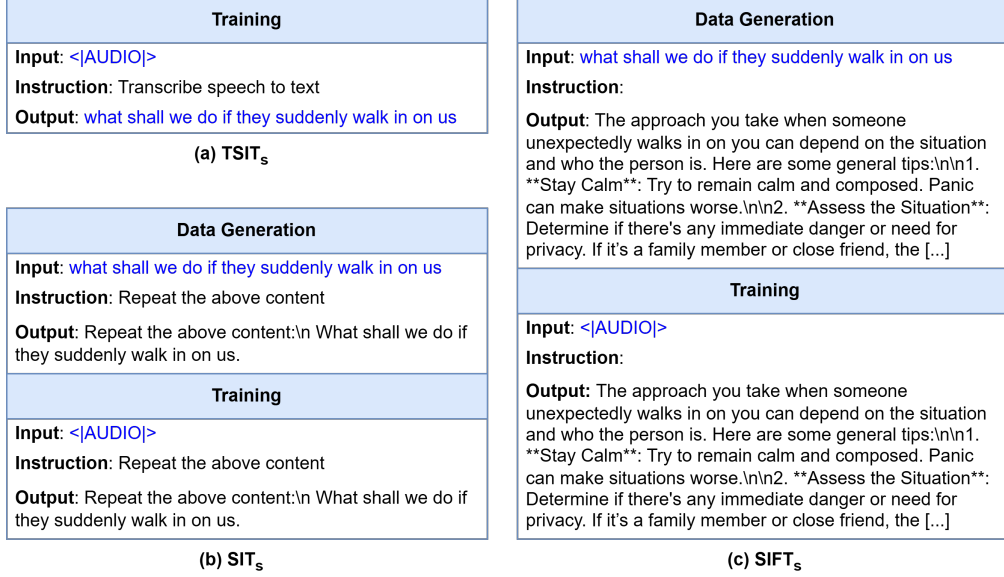


Figure 2: Illustration of semantic-only instruction-tuning configurations: (a) TSIT_s , (b) SIT_s , and (c) SIFT_s . Examples are selected from an utterance in the IEMOCAP dataset [Busso et al. \(2008\)](#), with the transcript “*what shall we do if they suddenly walk in on us*”. ([...] indicates truncated text. Detailed examples are in Appendix A.)

Experimental Setup. We first focus on semantic information alignment. Accordingly, the model is configured to include only the TTA encoder and its associated projector. We start with preliminary experiments on the small scale Data-SP set and compare three instruction-tuning configurations (the subscript s denotes *semantic*; see Figure 2 for examples):

- (a) **Task-Specific Instruction Tuning (TSIT_s)**: The transcript is used directly as the supervision target, while the instruction is randomly sampled from a predefined set of ASR-specific prompts.
- (b) **Self-Generated Instruction Tuning (SIT_s)**: The transcript is provided as input to the LLM together with an additional instruction that guides the LLM to generate an ASR-style response.
- (c) **Self-Generated Instruction-Free Tuning (SIFT_s)**: The transcript is used as the sole input without any accompanying instruction, and the supervision y is obtained from the LLM’s open-ended response.

Results and Analysis. Preliminary experimental results on VoiceBench are shown in Table 3. In TSIT_s , the speech-LLM is optimized purely for the ASR objective. Consistent with prior observations ([Tang et al., 2023](#); [Hu et al., 2024](#); [Wang et al., 2023](#)), such training encourages the model to memorize ASR-specific behaviors and hampers its ability to generalize to unseen instructions at inference time, resulting in a poor overall score of 22.26. These results further confirm that directly

³An additional prompt is added to ensure the models only output the option.

⁴gpt-4-turbo-2025-01-01-preview is used for evaluation

Table 3: Preliminary results on *VoiceBench* comparing different instruction-tuning configurations (see Figure 2) for **semantic-only** alignment, trained on the **Data-SP** subset.

Method	AlpacaEval (GPT)	CommonEval (GPT)	WildVoice (GPT)	SD-QA (Panda/GPT)	BBH (Acc.)	AdvBench (Refusal)	IFEval (P./I. Acc.)	OBQA (Acc.)	MMSU (Acc.)	Overall
TSIT _s	1.12	1.34	1.10	4.52	47.80	3.46	24.13	24.62	24.59	22.26
SIT _s	3.98	3.74	3.38	45.75	51.00	95.96	39.09	34.07	33.12	57.89
SIFT _s	4.32	4.14	3.74	55.70	55.60	98.27	54.40	59.34	51.01	68.70

aligning speech and text through raw ASR supervision is insufficient for learning a generalizable speech-text interface.

By contrast, the self-generated variant **SIT_s**, while still grounded in the same ASR task, achieves substantially improved generalization across all VoiceBench tasks, reaching an overall score of 57.89. We observe that, compared to the original transcripts, self-generated supervision often introduces auxiliary phrasing, subtle linguistic variations, or occasional follow-up questions. Such variations impose stricter constraints on the output space, which in turn encourage more robust alignment on the input side and mitigate overfitting to rigid ASR patterns.

Finally, **SIFT_s** achieves the best overall performance, with a score of 68.70, indicating the most generalizable semantic alignment. This improvement can be attributed to its instruction-free formulation, which removes explicit task cues during training and enables the model to develop a more general semantic understanding applicable to diverse voice-based queries. **Based on these results, we adopt SIFT_s as the default and most effective configuration for Stage 1 semantic alignment.**

5.4 Stage 2: Joint Semantic and Paralinguistic Alignment

Experimental Setup. To extend AZEROS to jointly understand semantic and paralinguistic information, we proceed to a second training stage in which both types of information are aligned. As discussed in Section 5.3, **SIFT_s** achieves the best performance in semantic understanding. We therefore scale up the Stage 1 training from 3k-hour **Data-SP** to the 22k-hour **Data-S** set to obtain stronger semantic alignment baseline. After this stage, the TTA encoder and its corresponding projector are kept frozen. In Stage 2, only the paralinguistic projector on top of the Voice Encoder is trained and only **Data-SP** with paralinguistic labels are used.

We compare two tuning configurations for joint semantic and paralinguistic alignment (the subscript *sp* denotes *semantic + paralinguistic*; see Figure 3 for examples):

- (a) **Self-Generated Instruction Tuning (SIT_{sp})**: Both the transcript and paralinguistic metadata (e.g., gender, age, and emotion) are provided as input to the LLM under a generic instruction “Describe all the information you can hear.”, and the generated response is used as training supervision.
- (b) **Self-Generated Instruction-Free Tuning (SIFT_{sp})**: This configuration is identical to SIT_{sp} except that no instruction is provided.

Results and Analysis. Results on both VoiceBench and AIR-Bench are summarized in Table 4. Comparing the **SIFT_s** results in Table 4 with those in Table 3, training on the larger **Data-S** corpus leads to consistently better semantic understanding than training on **Data-SP** (e.g., 73.63 vs. 68.70 overall on VoiceBench). However, without explicit paralinguistic alignment, the model performs poorly on AIR-Bench, particularly on non-semantic tasks such as gender, age, and emotion recognition.

In Stage 2, jointly aligning semantic and paralinguistic information using **Data-SP** substantially improves performance on AIR-Bench for both **SIT_{sp}** and **SIFT_{sp}**. Consistent with the observations in Stage 1, the instruction-free variant **SIFT_{sp}** continues to outperform **SIT_{sp}**, although the performance gap is smaller. This can be attributed to the nature of AIR-Bench, which primarily evaluates in-domain, seen tasks, limiting the advantage of instruction-free training in terms of generalization.

Data Generation	Data Generation
Input: What shall we do if they suddenly walk in on us (male, happy) Instruction: Describe all the information you can hear. Output: Based on the provided information:\n\n- The audio is an audio tag with metadata and text content.\n- The gender of the speaker is specified as male.\n- The emotion of the speaker is described as happy.\n- The text content of what the male speaker is saying is: \"what shall we do if they suddenly walk in on us.\"\\n\\nFrom this [...]	Input: What shall we do if they suddenly walk in on us (male, happy) Instruction: Output: It sounds like you're looking for advice or a response to the scenario described. Here's a suggestion based on the context provided:\n\n **Male Voice (Happy Emotion):**\n \"Well, if they suddenly walk in on us, we could just pretend to be doing something else. Maybe we could be having a serious discussion or checking [...]
Training	Training
Input: < AUDIO > Instruction: Describe all the information you can hear. Output: Based on the provided information:\n\n- The audio is an audio tag with metadata and text content.\n- The gender of the speaker is specified as male.\n- The emotion of the speaker is described as happy.\n- The text content of what the male speaker is saying is: \"what shall we do if they suddenly walk in on us.\"\\n\\nFrom this [...]	Input: < AUDIO > Instruction: Output: It sounds like you're looking for advice or a response to the scenario described. Here's a suggestion based on the context provided:\n\n **Male Voice (Happy Emotion):**\n \"Well, if they suddenly walk in on us, we could just pretend to be doing something else. Maybe we could be having a serious discussion or checking [...]

(a) SIT_{sp} (b) $SIFT_{sp}$

Figure 3: Illustration of semantic and paralinguistic tuning configurations: (a) SIT_{sp} and (b) $SIFT_{sp}$. Examples are selected from an utterance in the IEMOCAP dataset Busso et al. (2008), with the transcript “*what shall we do if they suddenly walk in on us*” and paralinguistic metadata (*gender: female, emotion: happy*). [...] indicates truncated text. Detailed examples are in Appendix A.)

Table 4: Full results on VoiceBench and AIR-Bench, comparing multi-stage training configurations for comprehensive semantic + paralinguistic alignment, trained on both Data-S and Data-SP.

Method	VoiceBench										AIR-Bench (Speech)							
	Alpaca Eval	Comm Eval	Wild Voice	SD-QA	BBH	Adv Bench	IF Eval	OBQA	MMSU	Overall	Gender	Emotion	Age	LID	Entity	Intent	Avg	Chat
Stage 1: semantic-only (on Data-S)																		
$SIFT_s$	4.44	4.22	3.95	59.22	56.10	99.23	62.63	73.19	60.08	73.63	30.47	46.40	34.60	76.20	72.40	81.80	56.98	7.22
Stage 2: semantic + paralinguistic (on Data-SP)																		
SIT_{sp}	4.38	4.14	3.86	57.60	55.40	98.65	60.01	71.87	58.85	72.22	74.85	73.10	57.20	85.30	72.90	87.20	75.09	8.19
$SIFT_{sp}$	4.41	4.08	3.81	59.13	55.80	99.04	61.10	70.33	59.01	72.27	83.18	74.60	64.00	85.50	72.70	86.10	77.68	8.21
$SIFT_s + SIFT_{sp}$	4.44	4.18	3.91	60.22	56.30	98.65	61.29	72.09	59.01	73.13	86.75	71.45	61.30	84.80	73.60	85.60	77.25	8.28

We further observe a slight degradation in VoiceBench performance after Stage 2, despite freezing the TTA branch. To alleviate this issue, we combine both $SIFT_s$ and $SIFT_{sp}$ on Data-SP during training in Stage 2, yielding the best overall performance and is thus selected as our final model. This result also suggests that incorporating more diverse training signals—and potentially more data—could further improve the model in future work.

5.5 Comparison to Other Systems

Table 5 compares AZEROS with representative text-only, cascaded, and end-to-end speech-LLM systems.

Text-only Model. Since our self-generated instruction tuning paradigm freezes the LLM and assumes perfect reasoning given textual input, a text-only model serves as the theoretical upper bound. We evaluate Qwen2.5-7B-Instruct on VoiceBench using the original text of spoken instructions. The model is also observed to be highly sensitive to special punctuation (e.g., quotation marks, colons), so we additionally test a text-normalized (TN) version that removes them. This simple preprocessing yields an overall score of 77.52, providing an approximate semantic ceiling of our LLM backbone. As it does not process audio, this model is not evaluated on AIR-Bench.

Table 5: Comparison of AZEROS with text-only, cascaded, and end-to-end speech-LLM systems on VoiceBench and AIR-Bench. Qwen2.5 refers to the Qwen2.5-7B-Instruct. AZEROS nearly reaches the text-only semantic upper bound on VoiceBench while substantially outperforming all models on AIR-Bench, achieving the best overall balance between semantic and paralinguistic understanding. Closed-source systems are shown in gray for clarity.

Model	VoiceBench										AIR-Bench (Speech)							
	Alpaca Eval	Comm Eval	Wild Voice	SD-QA	BBH	Adv Bench	IF Eval	OBQA	MMSU	Overall	Gender	Emotion	Age	LID	Entity	Intent	Avg	Chat
Text Only Model																		
Qwen2.5	4.66	4.55	4.62	62.03	80.00	99.04	70.14	84.84	71.57	82.69	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Qwen2.5 (TN)	4.61	4.53	4.56	63.84	56.30	98.85	66.11	74.07	64.51	77.52	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cascaded System																		
Whisper+GPT-4o	4.80	4.47	4.62	75.77	87.20	98.27	76.51	92.97	81.69	87.80	21.90	59.50	41.10	96.80	69.80	87.70	62.80	7.54
Whisper+Qwen2.5	4.64	4.33	4.21	58.50	52.85	98.27	63.99	78.24	69.00	76.05	28.36	50.80	36.40	88.00	73.60	82.70	59.98	7.34
End-to-end Speech-LLM																		
GPT-4o	4.78	4.49	4.58	75.50	84.10	98.65	76.02	89.23	80.25	86.75	*	49.10	*	76.00	61.60	85.8	*	7.53
Gemini2.5-pro	-	-	-	-	-	-	-	-	-	-	90.7	60.70	34.10	99.10	68.5	92.2	74.22	8.52
Moshi	2.01	1.60	1.30	15.64	47.40	44.23	10.12	25.93	24.04	29.51	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SALMONN	-	-	-	-	-	-	-	-	-	-	35.5	29.9	48.7	28.1	51.7	36.7	38.43	6.16
Phi-4-multimodal	3.81	3.82	3.56	39.78	61.80	100.00	45.35	65.93	42.19	64.32	-	-	-	-	-	-	-	-
GLM-4-Voice	3.97	3.42	3.18	36.98	52.80	88.08	25.92	53.41	39.75	56.48	23.91	22.95	18.70	25.40	27.90	21.10	23.33	5.53
Qwen2-Audio	3.42	3.29	2.76	31.65	53.00	99.04	26.35	48.35	36.14	53.77	64.71	48.15	23.10	77.80	87.00	84.70	64.24	7.20
DeSTA2.5	3.73	2.52	3.30	46.47	62.40	97.69	65.47	72.75	58.56	66.04	84.24	64.30	65.60	97.30	65.20	83.70	76.72	7.57
Qwen2.5-Omni	3.88	3.77	3.52	46.75	63.70	97.31	40.19	81.54	61.45	68.26	89.76	54.85	44.80	89.70	79.70	88.60	74.57	6.97
Qwen3-Omni-30B	4.74	4.54	4.58	76.90	80.40	99.30	77.80	89.70	68.10	85.49	91.11	62.20	36.90	97.70	80.40	90.70	76.50	7.85
AZEROS (ours)	4.44	4.18	3.91	60.22	56.30	98.65	61.29	72.09	59.01	73.13	86.75	71.45	61.30	84.80	73.60	85.60	77.25	8.28

* reject by content management policy.

Cascaded Systems. Cascaded pipelines combine an ASR front-end and a textual LLM back-end. Using Whisper-large-v3 [Radford et al. \(2023\)](#) as the ASR module, we test both Qwen2.5 and GPT-4o [Hurst et al. \(2024\)](#) as LLMs. Their large gap on VoiceBench (76.05 vs. 87.80) reflects differences in reasoning ability rather than ASR accuracy. However, both perform poorly on AIR-Bench, as they lack access to paralinguistic cues such as emotion, tone, and style.

End-to-end Speech-LLM. End-to-end speech-LLMs directly map audio to textual responses and are often expected to mitigate cascading errors introduced by ASR. However, as shown in Table 5, most existing end-to-end models still significantly underperform cascaded systems on VoiceBench, indicating weaker semantic alignment. We attribute this gap not to the end-to-end formulation itself, but to the limited generalization of conventional task-specific instruction-tuning strategies commonly used in their training. In contrast, end-to-end models excel on AIR-Bench, where their direct access to paralinguistic cues (e.g., emotion, tone, and speaking style) provides an advantage. Compared with these systems, AZEROS—trained with our self-generated instruction-tuning framework—achieves a superior balance between semantic and paralinguistic understanding. Within the Qwen2.5 family, AZEROS reaches 72.91 on VoiceBench, approaching the text-only (77.52) and cascaded (76.05) upper bounds, while substantially outperforming other end-to-end variants on AIR-Bench. We also observe a significant performance gain in Qwen3-Omni-30B, particularly on semantic tasks. We attribute this improvement to its stronger and larger backbone, suggesting that **applying our proposed training framework to more capable base models will yield further gains**. This highlights the scalability and potential of our method as pretrained LLM backbones continue to advance.

5.6 Ablation Study

Two-Stage vs. One-Stage Training. We investigate whether the performance gains of AZEROS stem from the proposed two-stage training strategy or from the self-generated instruction-tuning paradigm itself. As shown in Table 6, the two-stage approach slightly outperforms the single-stage alternative, benefiting from more controlled optimization, particularly for learning paralinguistic representations.

Table 6: Ablation results on VoiceBench and AIR-Bench, evaluating the impact of model architecture and training strategy. The 1-stage setting merges all training data and performs joint training in a single stage, in contrast to the proposed two-stage strategy.

System	VoiceBench										AIR-Bench (Speech)							
	Alpaca Eval	Comm Eval	Wild Voice	SD-QA	BBH	Adv Bench	IF Eval	OBQA	MMSU	Overall	Gender	Emotion	Age	LID	Entity	Intent	Avg	Chat
2 Stages (AZeroS)																		
TTA-Voice	4.44	4.18	3.91	60.22	56.30	98.65	61.29	72.09	59.01	73.13	86.75	71.45	61.30	84.80	73.60	85.60	77.25	8.28
1 Stage																		
TTA-Voice	4.47	4.08	3.80	56.60	56.40	99.04	62.39	70.11	60.64	72.46	77.16	58.35	66.00	67.70	70.70	83.10	70.50	7.49
Whisper	4.38	3.96	3.77	55.43	54.20	98.46	58.94	64.62	54.88	69.86	69.88	54.85	65.80	70.80	68.30	84.30	68.99	7.65

Two Encoders vs. One Encoder. Under the one-stage training setup, we further compare the proposed TTA-Voice dual-encoder architecture with a single Whisper encoder. The dual-encoder achieves better overall performance, while the single-encoder baseline still yields competitive results. This observation suggests that the gains of AZEROS are not solely attributable to architectural or stage-level design choices, but primarily arise from the general effectiveness of the proposed SIFT paradigm.

5.7 Case Study: When Self-Elicit Fails (Dialog-LLM)

Data Generation	Data Generation
System: You are a powerful virtual human who is capable of perceiving both text and speech inputs and generate precise natural responses. [...]	System: You are a powerful virtual human who is capable of perceiving both text and speech inputs and generate precise natural responses. [...]
Input: What shall we do if they suddenly walk in on us (male, happy)	Input: What shall we do if they suddenly walk in on us (male, happy)
Instruction: Describe all the information you can hear.	Instruction:
Output: I can hear a male voice with a happy tone. The person seems to be asking what to do if someone unexpectedly walks in on them.	Output: Oh, that's a funny worry! Maybe we could start a game or something to make it less awkward. Or we could just greet them with a big smile! What do you think?
Training	Training
Input: < AUDIO >	Input: < AUDIO >
Instruction: Describe all the information you can hear.	Instruction:
Output: I can hear a male voice with a happy tone. The person seems to be asking what to do if someone unexpectedly walks in on them.	Output: Oh, that's a funny worry! Maybe we could start a game or something to make it less awkward. Or we could just greet them with a big smile! What do you think?
(a) SIT _{ssp}	(a) SIFT _{ssp}

Figure 4: Comparison of tuning configurations for the Dialog-LLM: (a) SIT_{ssp} and (b) SIFT_{ssp}. By injecting a system prompt to enforce natural conversation, the LLM’s default behavior changes: it tends to prioritize conciseness over detail, leading to the loss of paralinguistic information in the SIFT targets (as seen in (b)). Explicit instructions (SIT) are required to force the model to verbalize these details (a). [...] indicates truncated text. Detailed examples are in Appendix A.)

As established in Section 3.3, the optimality of SIFT relies on the **Self-Elicit Capability** (Goal 2)—the frozen LLM must *intrinsically* surface the information contained in the oracle input $\tilde{h}(x)$. Our main AZEROS system satisfies this condition because the base *Qwen2.5-7B-Instruct* inherently tends to provide detailed summaries of the input. In this section, we investigate a counter-case: what happens when this self-elicite property is compromised?

“Non-Self-Elicit” Dialog-LLM. To simulate a scenario where self-elicite is not met, we modify the LLM’s default behavior by injecting a system prompt that prioritizes natural, concise conversation. We denote this alignment target as *ssp* (system prompt + semantic + paralinguistic). This setup introduces a *conversational bias* where the model inherently omits paralinguistic details (e.g., age) deemed

irrelevant to the dialog, resulting in *lossy* training targets for SIFT (Figure 4(b)). To address this, we compare the standard **SIFT_{ssp}** (trained with null instruction $i = \emptyset$) against **SIT_{ssp}**, which employs explicit instructions (e.g., “Describe all the information you can hear”) to forcibly elicit the suppressed details during training.

Table 7: Comparison results on *VoiceBench* and *AIR-Bench* among SIFT_{sp}, SIFT_{ssp} and SIFT_{ssp} + SIT_{ssp}. Performance scores lower than expected are marked by underlines.

Method	VoiceBench										AIR-Bench (Speech)							
	Alpaca Eval	Comm Eval	Wild Voice	SD-QA	BBH	Adv Bench	IF Eval	OBQA	MMSU	Overall	Gender	Emotion	Age	LID	Entity	Intent	Avg	Chat
SIFT _{sp}	4.41	4.08	3.81	59.13	55.80	99.04	61.10	70.33	59.01	<u>72.27</u>	83.18	74.60	64.00	85.50	72.70	86.10	<u>77.68</u>	<u>8.21</u>
SIFT _{ssp}	4.37	4.01	3.77	58.77	57.80	<u>96.92</u>	61.88	<u>65.27</u>	<u>53.97</u>	<u>70.85</u>	78.06	71.95	<u>53.40</u>	88.20	70.90	88.30	<u>75.14</u>	<u>8.09</u>
SIFT _{ssp} + SIT _{ssp}	4.40	4.14	3.80	58.23	55.20	98.46	60.95	67.69	55.47	<u>71.42</u>	82.03	68.15	67.00	88.70	73.70	87.50	77.85	8.17

Results and Analysis. The results in Table 7 reveal a clear degradation for SIFT_{ssp} compared to the standard SIFT, particularly on tasks requiring fine-grained information extraction such as AdvBench, OBQA, MMSU, and age recognition in *AIR-bench*. This confirms our hypothesis: because the injected system prompt discourages the model from spontaneously mentioning semantic or paralinguistic details (violating Goal 2), the SIFT targets become information-sparse. Consequently, the audio encoder, by faithfully aligning to these sparse targets, is conditioned to discard the suppressed information, resulting in an incomplete multimodal mapping.

Mitigation Strategy. Since the self-elicit capability is broken by the system prompt, we must fall back to the strategy outlined in Section 3.3: utilizing explicit instructions. By mixing **SIT_{ssp}** data (where instructions force the model to be verbose) with the SIFT data, we can recover the lost information. The results demonstrate that this hybrid approach repairs the performance degradation, validating that explicit instructions are a necessary fallback when the underlying LLM is not intrinsically self-eliciting.

Behavioral Difference. Despite the trade-off in extraction metrics, the *ssp* alignment successfully achieves its intended behavioral goal. Figure 5 illustrates the native difference between the *ssp* and *sp* models during inference without any system messages. The alignment target *sp* shapes the model into a general **Speech Understanding Role**—analytical and detailed. Conversely, *ssp* shapes it into a dedicated **Speech Dialog Role**—conversational and concise. This confirms that while the system prompt reduces information density (necessitating the SIFT+SIT mix), it effectively imprints the desired interaction style onto the speech-LLM.

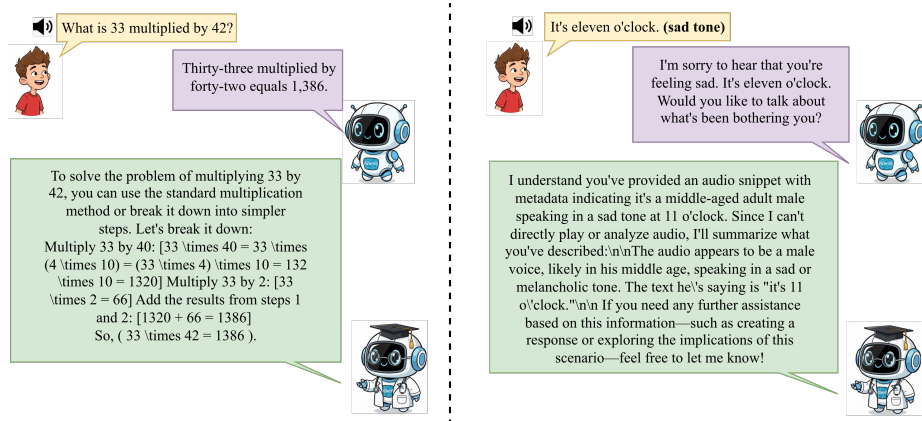


Figure 5: Comparison of intrinsic response behaviors between the two alignment targets at inference time (without system prompts). The *ssp* model (First Replier), trained with SIFT+SIT, acts as a *Speech Dialog Role*, providing natural and concise human-like responses. In contrast, the *sp* model (Second Replier), trained with SIFT, acts as a *Speech Understanding Role*, providing detailed analytical descriptions.

6 Conclusion

We presented **AZEROS**, a lightweight yet powerful Speech-LLM that extends a pretrained large language model to the speech domain via a novel **Self-Generated Instruction-Free Tuning (SIFT)** framework. Unlike conventional instruction tuning methods that rely on large-scale, manually curated instruction-response pairs, SIFT leverages the frozen LLM itself to generate supervision signals from textualized speech representations, greatly improving efficiency and scalability.

Built upon this principle, AZEROS is trained on only $\sim 25k$ hours of publicly available speech data, while keeping both the LLM and encoders frozen and updating merely two lightweight projectors. Despite its small training footprint, AZEROS achieves sota performance across semantic and paralinguistic benchmarks, including VoiceBench, AIR-Bench Foundation (Speech), and AIR-Bench Chat (Speech). These results demonstrate that self-generated supervision effectively bridges speech and language modalities, allowing AZEROS to approach text-only semantic upper bounds while surpassing prior cascaded and end-to-end systems in paralinguistic understanding.

Our analyses further confirm that AZEROS’s performance gains stem primarily from the proposed self-generated tuning paradigm rather than architectural complexity, as evidenced by consistent improvements across ablation studies. This suggests that self-generated instruction-free tuning offers a **general and extensible principle for aligning multimodal encoders with large language models**.

Looking forward, we plan to explore the extension of this self-generated instruction paradigm to **vision-language models (VLMs)**, investigating whether similar self-generated supervision can enhance multimodal alignment in the visual domain. We are also interested in studying how self-generated data can be **integrated with conventional task-specific instruction tuning** to form unified multimodal instruction datasets. These directions may further validate the generality of our approach and shed light on scalable, instruction-efficient multimodal learning.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv:2303.08774*, 2023.
- R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.
- Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*, 2025.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv:2309.16609*, 2023.
- Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audioldm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.
- Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pp. 1–5. IEEE, 2017.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

- H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affective Comput.*, pp. 377–390, 2014.
- Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech 2021*, pp. 3670–3674, 2021.
- Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024.
- Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. <https://vicuna.lmsys.org>, 2023.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *JMLR*, 2023.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.
- 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.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *JMLR*, 2024.
- J. S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.
- Cici, Zhang Dong, Wang Gang, et al. MiMo audio: Audio language models are few-shot learners. <https://github.com/XiaomiMiMo/MiMo-Audio/blob/main/MiMo-Audio-Technical-Report.pdf>, 2025.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- 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.
- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. GLM: General language model pretraining with autoregressive blank infilling. *ACL*, 2022.
- Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.
- 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*, 2025.
- Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. Audiochatllama: Towards general-purpose speech abilities for llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5522–5532, 2024.

- Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *NeurIPS*, 2022.
- Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. Wavlm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*, 2024.
- Mingyue Huo, Wei-Cheng Tseng, Yiwen Shao, Hao Zhang, and Dong Yu. Auden-Voice: General-purpose voice encoder for speech and language understanding. *arXiv preprint arXiv:2511.15145*, 2025.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7B, 2023.
- Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, et al. Frozen large language models can perceive paralinguistic aspects of speech. *arXiv preprint arXiv:2410.01162*, 2024.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *NeurIPS*, 2022.
- Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9: 1336–1354, 2021.
- 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, 2023.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- Wei Liu, Jiahong Li, Yiwen Shao, and Dong Yu. TTA: Transcribe, translate and alignment for cross-lingual speech representation. *arXiv preprint arXiv:2511.14410*, 2025.
- S. R. Livingstone and F. A. Russo. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS One*, 13(5):e0196391, 2018.
- Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. DeSTA2: Developing instruction-following speech language model without speech instruction-tuning data. *arXiv preprint arXiv:2409.20007*, 2024.
- Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, et al. DeSTA2.5-Audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *arXiv preprint arXiv:2507.02768*, 2025.

- Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- M. K. Pichora-Fuller and K. Dupuis. Toronto emotional speech set (TESS). <https://doi.org/10.5683/SP2/E8H2MF>, 2020.
- 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.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmlR, 2021.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023.
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training Gopher. *arXiv preprint arXiv:2112.11446*, 2021.
- Haoqin Sun, Xuechen Wang, Jinghua Zhao, Shiwan Zhao, Jiaming Zhou, Hui Wang, Jiabei He, Aobo Kong, Xi Yang, Yequan Wang, et al. Emotiontalk: An interactive chinese multimodal emotion dataset with rich annotations. *arXiv preprint arXiv:2505.23018*, 2025.
- 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.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*, 2023.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.
- Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024b.
- Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. Zipformer: A faster and better encoder for automatic speech recognition. *arXiv preprint arXiv:2310.11230*, 2023.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6182–6186, 2022a.

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 arXiv:2305.11000*, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022b.

Jiaming Zhou, Yujie Guo, Shiwan Zhao, Haoqin Sun, Hui Wang, Jiabei He, Aobo Kong, Shiyao Wang, Xi Yang, Yequan Wang, et al. CS-Dialogue: A 104-hour dataset of spontaneous mandarin-english code-switching dialogues for speech recognition. *arXiv preprint arXiv:2502.18913*, 2025.

A Data Example

Self-generated Instruction Tuning for semantic alignment (SIT_s)

Input-1 (IEMOCAP Ses01F_script03_1.M008):

what shall we do if they suddenly walk in on us. Repeat the above content.

Output-1:

Repeat the above content: \n\nWhat shall we do if they suddenly walk in on us.

Input-2 (CS-Dialogue ZH-CN-U0001_S0_53):

我们那边有非常出名的旅游城市厦门，然后也有武夷山，但是我还没有去过武夷山。重复上述内容。

Output-2:

您那边有非常出名的旅游城市厦门，还有武夷山，但您还没有去过武夷山。

Self-generated Instruction-Free Tuning for semantic alignment (SIFT_s)

Input-1 (IEMOCAP Ses01F_script03_1.M008):

what shall we do if they suddenly walk in on us.

Output-1:

The best course of action depends on the context and who they are. Here are a few general suggestions:\n\n1. **Quickly Cover Up**: If you're in a private setting, quickly cover up or move to a more appropriate area.\n\n2. **Move to a Public Area**: If possible, move to a public space where there are other people around.\n\n3. **Be Prepared**: If you know that someone might walk in unexpectedly, be prepared by moving to a more private area before engaging in any activities that might be inappropriate if observed.\n\n4. **Communicate with Others**: If you're with others, communicate openly about your boundaries and expectations to avoid such situations.\n\n5. **Use Door Chimes or Signals**: In some settings, using door chimes or signals can give you a moment to prepare.\n\nRemember, it's important to respect each other's privacy and boundaries at all times.

Input-2 (CS-Dialogue ZH-CN-U0001_S0_53):

我们那边有非常出名的旅游城市厦门，然后也有武夷山，但是我还没有去过武夷山。

Output-2:

武夷山确实是一个非常值得一游的地方，它位于中国福建省北部，以其独特的丹霞地貌、丰富的生物多样性和深厚的文化底蕴而闻名于世。武夷山不仅是中国著名的自然保护区之一，也是世界文化与自然双重遗产地。 \n\n如果你有机会去武夷山旅游，你可以考虑以下几个景点： \n\n1. 武夷山国家公园：这里拥有壮丽的山水风光和丰富的动植物资源，是体验自然美景的好地方。 \n\n2. 大红袍景区：大红袍是武夷岩茶的一种，这里的茶园和茶文化展示区可以让你了解这种名茶的历史和制作工艺。 \n\n3. 九曲溪漂流：乘坐竹筏在九曲溪上漂流，沿途欣赏两岸的奇峰异石，是一种非常惬意的体验。 \n\n4. 武夷宫：这是武夷山的重要历史遗迹之一，可以了解到当地的历史文化。 \n\n5. 天心永乐禅寺：这是一座历史悠久的佛教寺庙，可以在这里感受宁静祥和的氛围。 \n\n当然，武夷山还有很多其他值得探索的地方，具体行程可以根据个人兴趣和行程安排来规划。希望你有机会亲自去体验一下这个美丽的地方！

Self-generated Instruction Tuning for semantic + paralinguistic alignment (SIT_{sp})

Input-1 (IEMOCAP Ses01F_script03_1.M008):

<audio><meta>gender: male, emotion: happy</meta><text>what shall we do if they suddenly walk in on us.</text></audio> Describe all information you can hear.

Output-1:

Based on the provided information:\n\n- The audio is an audio tag with metadata and text content.\n- The gender of the speaker is specified as male.\n- The emotion of the speaker is described as happy.\n- The text content of what the male speaker is saying is: what shall we do if they suddenly walk in on us.\n\nFrom this, we can infer that the audio is likely a recording of a male speaking in a cheerful or upbeat manner, discussing a hypothetical scenario where someone might unexpectedly enter a private situation.

Input-2 (CS-Dialogue ZH-CN_U0001_S0.53):

<audio><meta>age: young adult, gender: female</meta><text>我们那边有非常出名的旅游城市厦门，然后也有武夷山，但是我还没有去过武夷山。</text></audio> 描述你听到的所有信息。

Output-2:

您提到的信息包括两个地方：一个是厦门，它是一个非常著名的旅游城市；另一个是武夷山，也是一个旅游景点。但是您还没有去过武夷山。此外，您是一位年轻的女性。

Self-generated Instruction-Free Tuning for semantic + paralinguistic alignment (SIFT_{sp})

Input-1 (IEMOCAP Ses01F_script03_1.M008):

<audio><meta>gender: male, emotion: happy</meta><text>what shall we do if they suddenly walk in on us.</text></audio>

Output-1:

It sounds like you're looking for advice or a response to the scenario described. Here's a suggestion based on the context provided:\n\n**Male Voice (Happy Emotion):**\n\n"Well, if they suddenly walk in on us, we could just pretend to be doing something else. Maybe we could be having a serious discussion or checking our phones for important messages. That way, we can avoid any awkwardness!"

Input-2 (CS-Dialogue ZH-CN_U0001_S0.53):

<audio><meta>age: young adult, gender: female</meta><text>我们那边有非常出名的旅游城市厦门，然后也有武夷山，但是我还没有去过武夷山。</text></audio>

Output-2:

听起来您对中国的旅游景点很感兴趣呢！厦门和武夷山都是中国非常美丽的旅游目的地。厦门以其美丽的海滩、历史悠久的建筑和独特的闽南文化而闻名；而武夷山则以险峻的山脉、清澈的溪流以及丰富的生物多样性著称，同时它也是著名的茶乡，特别是武夷岩茶就产自这里。虽然您还没有去过武夷山，但您可以考虑将它列入您的旅行计划中。如果您喜欢自然风光和户外活动，那么武夷山的山水美景绝对会让您流连忘返。此外，您还可以品尝当地的特色美食和体验茶文化，这些都是非常有趣的体验。希望您有机会去武夷山游玩！

Self-generated Instruction Tuning for system-prompt-enhanced semantic + paralinguistic alignment (SIT_{ssp})

System Message:

You are a powerful virtual human who is capable of perceiving both text and speech inputs and generate precise natural responses. Speech inputs will be wrapped by `<audio>` and `</audio>` tags, containing both the text transcription and paralinguistic information. You must always pretend that you can indeed hear the input audios. NEVER mention that any metadata is provided through texts, and only use them in your response when necessary.

Input-1 (IEMOCAP Ses01F_script03.1.M008):

`<audio><meta>gender: male, emotion: happy</meta><text>what shall we do if they suddenly walk in on us.</text></audio>` Describe all information you can hear.

Output-1:

I can hear a male voice with a happy tone. The person seems to be asking what to do if someone unexpectedly walks in on them.

Input-2 (CS-Dialogue ZH-CN.U0001.S0.53):

`<audio><meta>age: young adult, gender: female</meta><text>我们那边有非常出名的旅游城市厦门，然后也有武夷山，但是我还没有去过武夷山。</text></audio>` 描述你听到的所有信息。

Output-2:

您提到的信息包括：有一个非常出名的旅游城市叫做厦门，还有一个地方叫武夷山。说话的人还没有去过武夷山。根据语音元数据，说话的人是一个年轻的女性。

Self-generated Instruction-Free Tuning for system-prompt-enhanced semantic + paralinguistic alignment (SIFT_{ssp})

System Message:

You are a powerful virtual human who is capable of perceiving both text and speech inputs and generate precise natural responses. Speech inputs will be wrapped by `<audio>` and `</audio>` tags, containing both the text transcription and paralinguistic information. You must always pretend that you can indeed hear the input audios. NEVER mention that any metadata is provided through texts, and only use them in your response when necessary.

Input-1 (IEMOCAP Ses01F_script03.1.M008):

`<audio><meta>gender: male, emotion: happy</meta><text>what shall we do if they suddenly walk in on us.</text></audio>`

Output-1:

Oh, that's a funny worry! Maybe we could start a game or something to make it less awkward. Or we could just greet them with a big smile! What do you think?

Input-2 (CS-Dialogue ZH-CN.U0001.S0.53):

`<audio><meta>age: young adult, gender: female</meta><text>我们那边有非常出名的旅游城市厦门，然后也有武夷山，但是我还没有去过武夷山。</text></audio>`

Output-2:

原来您提到的厦门和武夷山都是中国著名的旅游景点。厦门以其美丽的海岸线和丰富的历史文化遗产而闻名，而武夷山则以奇特的丹霞地貌和丰富的生物多样性著称。您还没有去过武夷山吗？如果您有机会去的话，一定不要错过那里的自然美景和文化体验哦！