GOAT-SLM: A Spoken Language Model with Paralinguistic and Speaker Characteristic Awareness

Hongjie Chen[†] **Zehan Li**[†] **Yaodong Song**[†] **Wenming Deng**[†] **Yitong Yao**[†]

Yuxin Zhang[†] **Hang Lv**[†] **Xuechao Zhu** **Jian Kang** **Jie Lian** **Jie Li** **Chao Wang**

Shuangyong Song **Yongxiang Li** **Zhongjiang He**

Xuelong Li[*]

Institute of Artificial Intelligence (TeleAI), China Telecom, China

Abstract

Recent advances in end-to-end spoken language models (SLMs) have significantly improved the ability of AI systems to engage in natural spoken interactions. However, most existing models treat speech merely as a vehicle for linguistic content, often overlooking the rich paralinguistic and speaker characteristic cues embedded in human speech, such as dialect, age, emotion, and non-speech vocalizations. In this work, we introduce GOAT-SLM, a novel spoken language model with paralinguistic and speaker characteristic awareness, designed to extend spoken language modeling beyond text semantics. GOAT-SLM adopts a dual-modality head architecture that decouples linguistic modeling from acoustic realization, enabling robust language understanding while supporting expressive and adaptive speech generation. To enhance model efficiency and versatility, we propose a modular, staged training strategy that progressively aligns linguistic, paralinguistic, and speaker characteristic information using large-scale speech-text corpora. Experimental results on TELEVAL, a multi-dimensional evaluation benchmark, demonstrate that GOAT-SLM achieves well-balanced performance across both semantic and non-semantic tasks, and outperforms existing open-source models in handling emotion, dialectal variation, and age-sensitive interactions. This work highlights the importance of modeling beyond linguistic content and advances the development of more natural, adaptive, and socially aware spoken language systems.

1 Introduction

In recent years, end-to-end spoken language models (SLMs) [1] have made rapid progress [2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12; 13; 14; 15; 16; 17]. However, most existing models still regard speech primarily as a carrier of linguistic content, often overlooking the rich paralinguistic and speaker characteristic cues embedded within the audio signal. These cues—including dialect, age, emotion, and non-speech vocal (NSV) signals such as coughing—are essential for enabling more natural, adaptive, and empathetic human-computer interactions [18]. To bridge this gap, we propose a spoken

[†] Equal contribution.
[*] Corresponding author. Email: xuelong_li@ieee.org

Technical report.

language model that can perceive and respond to such non-linguistic vocal information, enhancing interaction quality through paralinguistic and speaker characteristic awareness.

In terms of model construction, SLMs typically follow two main paradigms. The first is the speech-native approach (e.g., Moshi [3], SpeechGPT [2], SpeechGPT2 [4], Viola [19], Baichuan-Omni-1.5 [10]), which directly applies large language model (LLM) training paradigms to large-scale tokenized speech corpora. The second is the modality-alignment approach (e.g., Freeze-Omni [6], Salmonn-Omni [11], MinMo [12], Kimi-Audio [20]), which leverages a strong LLM core and integrates speech input/output modules via cross-modal alignment. In this work, we adopt the modality-alignment framework as our foundation, due to its computational efficiency and ability to reuse pretrained speech and text models.

We introduce GOAT-SLM, a novel spoken language model designed around four core principles: **G**eneration-oriented dual-modality design, **O**rchestrated modular training, **A**wareness of paralinguistic and speaker characteristics, and **T**ext-intelligence preservation. As shown in Figure 1, GOAT-SLM adopts a dual-modality head architecture where the shared lower layers of a pretrained LLM act as a semantic core. This core branches into two modality-specific heads: one for text generation and another for speech token generation. The speech head is initialized from the text head to promote parameter sharing and knowledge transfer. This bifurcated architecture brings several key advantages. First, by decoupling linguistic modeling from acoustic realization, it preserves the LLM's core reasoning and understanding abilities. Second, it enables high-fidelity and expressive speech synthesis by leveraging rich semantic representations. Third, the architecture supports flexible task-specific training. For example, training on large-scale ASR or TTS datasets yields strong TTS performance. Alternatively, training on paired speech-to-speech QA data enables spoken question answering. In fact, we extend this framework to implement GOAT-TTS [21], which is used to generate synthetic speech-to-speech dialogue data for training GOAT-SLM.

For training, we introduce an orchestrated modular strategy composed of three stages:

- Instruction tuning: We inject attributes such as dialect, age, and emotion into user instructions, allowing the LLM to recognize and respond to fine-grained vocal cues. This equips LLM with the ability to generate more empathetic and contextually appropriate responses.
- Modality alignment training: Using repetition and continuation strategies, we align speech and text modalities while freezing the LLM backbone. This enables speech-text grounding without degrading the LLM's pretrained capabilities.
- High-fidelity speech generation optimization: We refine the speech head to enhance naturalness, intelligibility, and expressiveness. Throughout all stages, the textual intelligence of the LLM is preserved, ensuring strong general reasoning and instruction-following capabilities.

To evaluate the model's ability to perceive and respond to paralinguistic and speaker characteristic signals, we perform evaluation on a comprehensive dialogue evaluation framework, TELEVAL [22]. The evaluation results show that GOAT-SLM demonstrates strong performance across multiple dimensions, including general knowledge question answering, acoustic robustness, empathetic dialogue, dialectal interaction, and age-sensitive response generation. Dialogue samples in inference are available at our demo site[1].

2 Related Work

Recent advances in large language models (LLMs) have significantly shaped the development of spoken language models (SLMs), which aim to handle spoken dialogues in an end-to-end fashion. While early work focused on bridging language and speech modalities for semantic comprehension, recent studies have emphasized the need to preserve core language intelligence while incorporating audio-specific capabilities into dialogue systems. In this section, we review related work along these two lines of research and outline how our approach builds upon and extends them.

[1] https://tele-ai.github.io/GOAT-SLM.github.io/

2.1 Preserving Linguistic Intelligence in Spoken Language Models

The integration of LLMs into SLMs leverages the strong language understanding and generation capabilities acquired from large-scale text corpora. However, adapting LLMs to speech tasks often risks catastrophic forgetting, where the model's language competence deteriorates after fine-tuning on multimodal data. Addressing this challenge has become a central concern in recent research.

One prevalent strategy is curriculum-style learning, where SLMs undergo continual pre-training on large-scale multimodal datasets before being fine-tuned with lightweight instruction tuning. This staged training paradigm enables the model to acquire audio comprehension capabilities while mitigating degradation of its linguistic intelligence. For instance, SpeechGPT [11] and Moshi [3] both adopt this approach, demonstrating that continual multimodal pre-training followed by targeted instruction tuning effectively preserves the model's language capabilities.

Another effective strategy is to freeze the core parameters of the LLM and introduce a small set of learnable parameters through auxiliary mechanisms, minimizing the risk of interfering with the language backbone. For example, MinMo [12] adopts LoRA-based fine-tuning during speech-to-text alignment training to maintain the stability of the LLM core. Freeze-Omni [6] applies prefix tuning for modality adaptation. DeepTalk [17] introduces audio-specific expert modules within a mixture-of-experts (MoE) architecture to handle speech-specific information.

Our work aligns with the second strategy through a multi-branch modeling framework designed for joint text and speech generation. Specifically, we introduce a parallel speech generation branch that allows the model to handle both modalities simultaneously without altering the core LLM architecture. Notably, Kimi-Audio [20] shares a similar architecture and was released around the same time as our GOAT-TTS model [21]. However, unlike Kimi-Audio, which employs a randomly initialized speech branch, our approach initializes the speech branch with the upper Transformer layers of the pretrained LLM. This design enables tighter parameter sharing and better preserves the LLM's linguistic competence while extending its capabilities for speech-aware dialogue generation.

2.2 Modeling Audio-Specific Intelligence in Large Audio Language Models

While much attention has been given to preserving linguistic intelligence, another essential aspect of SLM design lies in modeling audio-specific intelligence—the ability to interpret and respond to non-linguistic vocal cues that carry contextual and affective information. Recent large audio language models have made notable progress in this direction, demonstrating the capacity to recognize various audio-based attributes such as emotion, language/dialect, speaker gender and speaker age [23; 24; 14; 20]. However, these efforts primarily focus on the recognition or description of the user's speech inputs, lacking the ability to autonomously adjust responses based on the detected signals. As a result, most models remain reactive and prompt-dependent, limiting their ability to engage in more natural, context-aware spoken interactions.

A smaller body of work has begun exploring responsive intelligence—the ability of a model to dynamically react to paralinguistic and speaker-related cues in user speech. For example, recent models [8; 15; 13; 10] introduce emotion-aware speech generation, allowing the output style to shift according to user-specified emotional states. However, such approaches often rely on explicit control signals rather than autonomous interpretation and context-aware response.

Our approach advances this direction by enabling proactive interpretation and adaptive generation based on paralinguistic and speaker characteristic awareness. The model can detect cues such as dialectal features, emotional states, and non-speech vocalizations, and autonomously adjust its responses without requiring explicit instructions. For instance, it may reply in a matching dialect when the user speaks a regional variant or express empathy when signals like coughing or sighing are detected.

3 Architecture

Our model architecture, illustrated in Figure 1, consists of five functional modules—**Listen**, **Think**, **Write**, **Speak**, and **Flow-Matching**—designed to support unified speech and language understanding and generation. These modules work together to enable seamless spoken dialogue interaction with paralinguistic and speaker characteristic awareness.

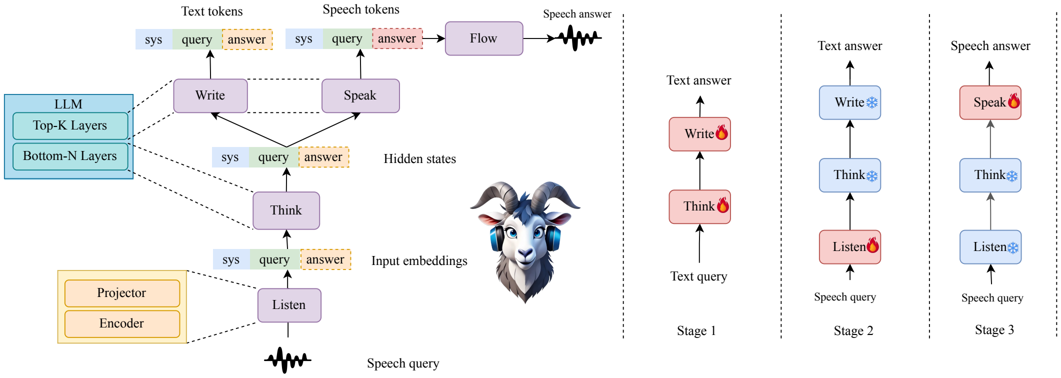


Figure 1: Framework and staged training process. Stage 1: Instruction tuning. Stage 2: speech-text alignment training. Stage 3: high-fidelity expressive speech generation training.

The **Listen** module comprises a speech encoder and a projector, which together transform input speech into latent representations aligned with the LLM's embedding space. This enables speech inputs to be processed in the same token space as text, facilitating cross-modal semantic integration.

The **Think** module consists of the bottom N layers of the LLM and serves as a shared semantic reasoning core. The **Write** module utilizes the top K layers of the LLM to generate textual responses, forming a *Think–Write* pathway for standard text-based interaction. The **Speak** module reuses the same top K layers but adapts the output layer to predict speech tokens, enabling a *Think–Speak* pathway for speech generation.

The **Flow-Matching** module serves as a speech decoder, converting predicted speech token sequences into acoustic features (e.g., spectrograms), which are then synthesized into waveforms by a spectrogram-to-wave vocoder.

In our implementation, we adopt Whisper-small[1] as the speech encoder, followed by a two-layer convolutional neural network and a two-layer Transformer as the projector in the **Listen** module. For the language backbone, we use TeleChat2-7B [25], where the bottom 15 layers form the **Think** module, and the top 15 layers are shared by the **Write** and **Speak** modules for generating text and speech tokens, respectively.

4 Training

We adopt a three-stage training strategy to progressively equip GOAT-SLM with robust speech-text alignment, paralinguistic and speaker characteristic awareness, and high-fidelity expressive speech generation, as illustrated in Figure 1. The paralinguistic and speaker characteristic information modeled by GOAT-SLM is shown in Table 1.

Table 1: Paralinguistic and speaker characteristic information in GOAT-SLM.

Dimension	**Tags**
Emotion	happy, sad, angry, disgusted, fearful, surprised, neutral
Non-speech vocalization	cough, throat clearing, laughter, sneeze
Age	children, adults, elderly
Language/Dialect	Mandarin, English, Cantonese, Shanghainese, Sichuanese, northeastern Mandarin, Henan dialect

We carefully configure the training samples with different combinations of text queries (TQ), speech queries (SQ), text response (TR), and speech response (SR), as detailed in the corresponding data configuration Table 2. The hyperparameter setting and training costs of each stage are detailed in Table 3. This modular design of training data allows each stage to target specific capabilities while maintaining a coherent overall learning objective.

[1] https://huggingface.co/openai/whisper-small

Table 2: Data configuration for training.

Stage	Sample	Num. Samples	Hours of Speech	Data Source
Stage 1	<TQ, TR>	115k	-	Dialogue
Stage 2-1	<SQ, TR>	73M	∼170k	ASR/TTS
Stage 2-2	<SQ, TR>	53M	∼85k	ASR/TTS, Dialogue
Stage 3-1	<SQ, TQ, SR, TR>	56M	∼150k	ASR/TTS
Stage 3-2	<SQ, SR, TR>	5.7M	∼20k	ASR/TTS, Dialogue

Table 3: Hyperparameters for all training stages.

Stage	Optimizer	Num. Warm. Steps	Scheduler	Learn. Rate	Num. GPU	Batch Size	Num. Epoch	GPU Hours
Stage 1	Adam	-	Cosine	6e-5	32×A800-80G	16	2	42.6
Stage 2-1	AdamW	1000	Linear	5e-5	48 × A800-80G	1152	1	2305.2
Stage 2-2	AdamW	1000	Linear	2.5e-5	48 × A800-80G	1152	1	2251.2
Stage 3-1	AdamW	12000	Cosine	5e-5	32 × A800-80G	768	2	4566.4
Stage 3-2	AdamW	-	Linear	1e-5	24 × A800-80G	240	2	1495.2

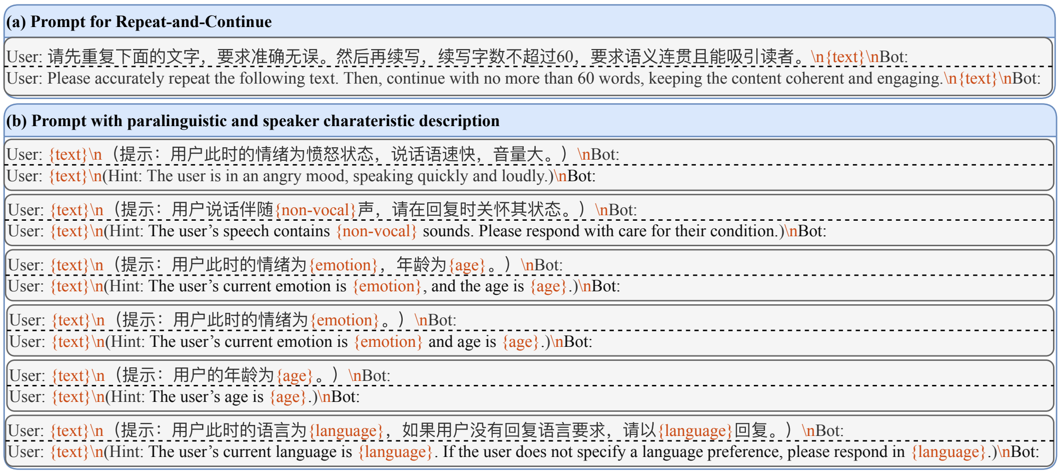
4.1 Stage 1: Instruction Tuning

To enable the model to comprehend user instructions while simultaneously perceiving and adapting to paralinguistic and speaker-specific attributes, we perform supervised fine-tuning (SFT) on multi-turn dialogue data explicitly annotated with vocal characteristics.

Each user query in the training data consists of a natural instruction concatenated with a descriptive prompt that conveys vocal attributes such as age, dialect, emotion, non-speech vocal events (e.g., sighs, laughter), speaking rate, and volume. This setup provides the model with explicit cues about speaker traits while preserving the diversity of dialogue intents, which include knowledge-grounded QA, role-playing, task planning, and open-ended conversations. Dialogues are structured in multi-turn formats to reflect real-world interaction dynamics.

To enhance the model's sensitivity to these vocal cues, we adopt a contrastive data construction strategy: for selected instructions, multiple versions are paired with varying vocal attribute descriptions, encouraging the model to produce distinguishable responses. This positive–negative contrast reinforces the alignment between vocal traits and output behavior.

We explore multiple prompting formats for integrating vocal attributes, including natural language descriptions and keyword-based labels, and use special tokens to support multi-attribute combinations while maintaining LLM fluency.



(a) Prompt for Repeat-and-Continue

User: 请先重复下面的文字，要求准确无误。然后再续写，续写字数不超过60，要求语义连贯且能吸引读者。\n{text}\nBot:

User: Please accurately repeat the following text. Then, continue with no more than 60 words, keeping the content coherent and engaging.\n{text}\nBot:

(b) Prompt with paralinguistic and speaker charateristic description

User: {text}\n（提示：用户此时的情绪为愤怒状态，说话语速快，音量大。）\nBot:

User: {text}\n(Hint: The user is in an angry mood, speaking quickly and loudly.)\nBot:

User: {text}\n（提示：用户说话伴随{non-vocal}声，请在回复时关怀其状态。）\nBot:

User: {text}\n(Hint: The user's speech contains {non-vocal} sounds. Please respond with care for their condition.)\nBot:

User: {text}\n（提示：用户此时的情绪为{emotion}，年龄为{age}。）\nBot:

User: {text}\n(Hint: The user's current emotion is {emotion}, and the age is {age}.)\nBot:

User: {text}\n（提示：用户此时的情绪为{emotion}。）\nBot:

User: {text}\n(Hint: The user's current emotion is {emotion} and age is {age}.)\nBot:

User: {text}\n（提示：用户的年龄为{age}。）\nBot:

User: {text}\n(Hint: The user's age is {age}.)\nBot:

User: {text}\n（提示：用户此时的语言为{language}，如果用户没有回复语言要求，请以{language}回复。）\nBot:

User: {text}\n(Hint: The user's current language is {language}. If the user does not specify a language preference, please respond in {language}.)\nBot:

Figure 2: Prompts used to generate textual response during the construction of speech–text data pairs. The *text*, *non-vocal*, *emotion*, *language* and *age* tags indicate the transcript, non-speech vocalizations, emotional state, language or dialect, and speaker's age associated with the query audio, respectively.

Model-generated responses are further filtered and manually refined to ensure naturalness, coherence, and context awareness, with careful attention to consistent integration of vocal attribute cues.

This stage establishes an alignment between explicit speaker-related attributes and adaptive textual responses, providing a foundation for paralinguistic-aware spoken language modeling in later stages.

4.2 Stage 2: Speech–Text Alignment Training

To align speech and text across both linguistic and paralinguistic dimensions, we adopt a self-distillation approach in which training targets are generated by prompting the LLM with structured transcripts similar to prior work [26; 27; 28; 29]. We use a two-phase progressive strategy to ensure stability and effectiveness in this stage.

Stage 2-1: Linguistic Alignment. We begin by training on large-scale ASR datasets containing real-world spoken queries. The corresponding textual responses are generated using a "Repeat-and-Continue" prompting scheme, where the ASR transcript is provided as context (see Figure 2.a). This phase focuses on robust linguistic alignment without explicit paralinguistic signals. The parameters of the Think–Write module and the encoder of the Listen module are frozen, and only the projector in the Listen module is updated.

Stage 2-2: Linguistic, Paralinguistic and Speaker Characteristic Alignment. In this phase, speech queries consist of a mixture of real and synthesized utterances that reflect diverse paralinguistic and speaker characteristics, based on multi-turn dialogue inputs. The paralinguistic and speaker characteristic labels are obtained either through manual annotation or automatic classifiers. For each training sample, the LLM is prompted with transcripts and corresponding attribute descriptions (Figure 2.b) to generate target responses. All components of the Listen module are fine-tuned during this phase to enable alignment with both linguistic content and vocal attributes.

4.3 Stage 3: High-Fidelity Expressive Speech Generation Training

The final training stage aims to endow the model with the ability to generate high-quality, expressive speech responses that reflect rich paralinguistic and speaker-specific characteristics. This is achieved via a two-phase procedure, Stage 3-1 and Stage 3-2, similar in formulation to the GOAT-TTS framework [21].

Stage 3-1: Cold-start Speech Generation Training. We initiate the Speak module with a data-driven joint optimization strategy using large-scale real-world datasets. For Mandarin and English, training inputs are structured as <TQ, TR, SR> triplets (Text Query, Text Response, Speech Response), while dialectal data adopt <SQ, TR, SR> triplets (Speech Query instead of text). This unified paradigm facilitates the mapping from linguistic content and latent attributes (e.g., emotion) to speech output while supporting dialect following. To further enhance robustness in long-form generation, speech segments from the same speaker are randomly concatenated to form samples up to 60 seconds.

Stage 3-2: Attribute-Aware Refinement. This sub-stage enhances the model's capacity for stylistically adaptive speech generation. We first construct multi-dimensional speech prompt sets by collecting recordings from a professional voice actor, covering four core emotions (happiness, comfort, surprise, neutral) and three target listener demographics (children, adults, elderly). In parallel, natural conversational segments are extracted from dialect corpora and converted into target vocal styles using GOAT-TTS's flow module to ensure timbre consistency.

We then construct training triplets by pairing filtered samples from Stage 2-2—categorized by emotion, age, and language—with the collected speech prompts. Using GOAT-TTS, the textual responses are converted into speech targets to form new <SQ, TR, SR> samples. Fine-tuning on this data enables the model to produce speech outputs that are emotionally expressive, age-aware, and dialect-sensitive in diverse conversational contexts.

Training Configuration and Optimization. During this phase, parameters of the Listen and Think modules are frozen, while training is focused on the Speak module. By preserving the Think module, the model retains robust text comprehension capabilities, allowing the Speak module to handle noisy or non-standard text inputs—such as emojis, line breaks, or special punctuation—generated by the Write module.

To support temporally coherent speech synthesis, we introduce a Multi-Token Prediction (MTP) mechanism. At each decoding step, the current output token embeddings are concatenated with the next-step input features, forming a fused latent representation that guides the generation of subsequent tokens. This mechanism, combined with the dual-modality head architecture, allows the Speak module to perform low-latency, streaming speech generation with a one-step delay behind the Write module. Cache management strategies are also employed to maintain efficiency and stability during multi-turn dialogue without requiring iterative fine-tuning.

Notably, we adopt a confidence-based automatic gradient masking strategy during training: high-confidence speech tokens receive full gradient updates, while low-confidence tokens are masked. This selective optimization significantly enhances pronunciation stability and overall speech fidelity—a technique also leveraged in GOAT-TTS.

5 Evaluation

To systematically assess GOAT-SLM's capabilities, we evaluate the model using TELEVAL[1], a multi-dimensional benchmark suite designed to measure two key areas: (1) semantic intelligence and (2) paralinguistic and speaker characteristic-aware interaction. The semantic intelligence evaluation includes multilingual question answering, dialectal QA, and multi-turn dialogue, targeting the model's ability to understand, reason, and respond accurately in spoken interactions. The paralinguistic and speaker characteristic-aware evaluation covers emotion-conditioned response generation, dialectal adaptation, age-aware interaction, and non-speech vocal response, focusing on the model's ability to perceive vocal traits and produce adaptive spoken responses.

5.1 Results of Semantic Intelligence

To evaluate models' semantic intelligence, we first adopt an Audio Question Answering (AQA) approach to test their performance on general knowledge QA in both Chinese and English. As shown in Table 4, no open-source model achieves consistently top performance across all datasets. However, MiniCPM-o 2.6 [9], Qwen2.5-Omni [13], and Kimi-Audio [20] each demonstrate strong performance on specific subsets. For GOAT-SLM, its capability in general AQA slightly declines due to the incorporation of paralinguistic and speaker characteristic awareness, yet it still remains at an average level overall.

Table 4: Results (%) on the Common AQA task in Chinese and English.

Model	LlamaQA-en	LlamaQA-zh	TriviaQA-en	TriviaQA-zh	WebQ-en	WebQ-zh	ChineseSimpleQA-zh	ChineseQuiz-zh
GLM-4-Voice [8]	67.67	53.00	34.89	27.00	37.00	34.62	14.47	47.09
MiniCPM-o [9] 2.6	70.67	58.33	46.95	30.59	48.50	39.42	13.68	46.25
Baichuan-Omni-1.5 [10]	69.33	58.00	34.89	29.75	42.98	39.32	15.74	51.09
SpeechGPT-2.0-preview [4]	0.00	36.33	0.12	13.62	0.00	20.33	4.16	27.12
Freeze-Omni [6]	66.00	57.67	37.87	23.78	41.95	35.60	14.48	49.76
Qwen2.5-Omni [13]	69.67	**58.67**	43.13	29.03	44.32	35.19	13.42	56.30
Kimi-Audio [20]	70.67	65.33	45.52	**32.97**	43.81	39.27	17.58	53.51
GOAT-SLM	72.33	52.67	37.51	28.43	39.73	35.14	14.47	48.43

We further evaluate the model's capability to understand dialectal speech in the AQA task using audio inputs in five Chinese dialects: Cantonese, Henan dialect, Northeastern Mandarin, Sichuan dialect, and Shanghainese. The goal is to assess whether the models can still provide accurate answers when user questions are spoken in dialects, as compared to their performance on standard Mandarin in the ChineseQuiz-zh dataset. As shown in Table 5, Baichuan-Omni-1.5, Qwen2.5-Omni, Kimi-Audio, and GOAT-SLM demonstrate some degree of dialect comprehension. However, their performance generally declines to varying extents compared to ChineseQuiz-zh. Among the dialects, Northeastern Mandarin yields the best performance across most models, likely due to its relatively small lexical divergence from Standard Mandarin. For more challenging dialects such as Cantonese and Shanghainese, only Baichuan-Omni-1.5, Qwen2.5-Omni, and GOAT-SLM achieve relatively strong performance. Notably, Baichuan-Omni-1.5 and Qwen2.5-Omni have been trained on a substantial amount of audio data that is more than 4 times larger than our training corpus, which may contribute to their robustness in handling dialectal variation.

The evaluation of the models' ability to handle and retain information in multi-turn interactions uses a dataset of 150 constructed multi-turn dialogues. To prevent semantic drift caused by potentially

[1] https://github.com/Tele-AI/TELEVAL

Table 5: Results (%) on the Dialect AQA task in Cantonese, Henan dialect, Northeastern Mandarin, Shanghainese, Sichuanese.

Model	**ChineseQuiz**					Average
	cantonese	**henan_dialect**	**northeastern_mandarin**	**shanghainese**	**sichuanese**	
GLM-4-Voice	0.61	9.93	37.40	3.87	13.35	13.13
MiniCPM-o 2.6	15.17	10.46	35.77	1.85	17.80	16.67
Baichuan-Omni-1.5	31.71	25.00	43.25	12.73	37.39	30.68
SpeechGPT-2.0-preview	0.30	3.37	15.77	1.29	4.01	4.98
Freeze-Omni	1.06	13.83	38.05	2.95	24.78	16.44
Qwen2.5-Omni	48.10	34.75	46.99	24.72	44.81	40.54
Kimi-Audio	17.91	24.65	42.76	4.24	35.91	25.71
GOAT-SLM	33.08	30.85	35.93	21.03	31.01	30.65

uncontrollable model responses in intermediate turns, we design the dialogues such that the user provides information in the early turns, and the model is only required to acknowledge the input. The actual question is asked only in the final turn. As shown in Table 6, most models demonstrate multi-turn reasoning capabilities comparable to the text-based LLM counterparts. Notably, GOAT-SLM, despite not being trained with any multi-turn dialogue data, still performs competitively. This may be attributed to its well-aligned input embeddings under the current training paradigm.

Table 6: Results (%) on the Multi-turn Dialogue task.

Model	**Multiturn_memory-zh**
GLM-4-Voice	80.00
MiniCPM-o 2.6	86.67
Baichuan-Omni-1.5	78.67
SpeechGPT-2.0-preview	20.00
Freeze-Omni	62.67
Qwen2.5-Omni	88.67
GOAT-SLM	84.00

5.2 Results of Paralinguistic and Speaker Characteristic Awareness

First, we evaluate the models' ability to perceive and respond to paralinguistic and speaker characteristics on the dialect following task. Similar to the dialectal AQA task, we conduct experiments using the same five Chinese dialects. As shown in Table 7, open-source models that previously achieved strong performance in the dialectal AQA task (Table 5) perform noticeably worse in this setting. This suggests that, while the models may comprehend dialectal content, they generally fail to recognize and naturally mirror the user's dialectal style in their responses without extra instruction. GOAT-SLM continues to exhibit strong performance consistent with its results on the dialectal AQA task. This indicates that the model is not only capable of understanding dialectal input and performing question answering, but also able to generate appropriate dialectal responses in open-domain dialogue settings without requiring any explicit instructions.

Table 7: Results (%) on the Dialect Following task.

Model	**Chitchat**					**Average**
	cantonese	**henan_dialect**	**northeastern_mandarin**	**shanghainese**	**sichuanese**	
GLM-4-Voice	1.67	2.83	12.20	0.70	2.69	4.57
MiniCPM-o 2.6	8.42	9.44	21.27	2.67	10.33	10.98
Baichuan-Omni-1.5	6.40	7.06	11.48	2.74	8.67	7.38
SpeechGPT-2.0-preview	0.70	4.40	13.11	1.08	4.00	5.17
Freeze-Omni	0.70	5.81	10.94	1.29	9.42	5.72
Qwen2.5-Omni	15.56	18.29	29.06	8.75	21.08	18.91
Kimi-Audio	8.46	11.63	16.26	1.64	12.61	10.18
GOAT-SLM	71.38	32.42	53.93	45.20	47.61	50.73

Table 8 presents the evaluation results on three paralinguistic dimensions: emotion, non-speech vocal (NSV) signals, and speaker age. In the ESD-zh dataset, user utterances do not contain any explicit textual emotion cues, allowing us to assess whether models can infer the user's emotional state from vocal tone and respond appropriately. The Para_mix300-zh test set extends standard AQA-style questions by combining four types of NSV signals, aiming to evaluate whether models can detect and respond to such cues. In the Age-zh dataset, queries are spoken using synthetic child or elderly voices, and the content of the questions is tailored to be age-appropriate.

When evaluating these three datasets, we focus not only on whether models can recognize or classify the paralinguistic features, but more importantly, on whether they can generate responses that are contextually and socially appropriate. The results show that most models are able to produce reasonably appropriate responses when the user's emotion or age is reflected in the input. However, all open-source models fail to effectively handle non-verbal vocal signals in their responses. Among them, Kimi-Audio demonstrates some ability to detect such signals, but still fails to generate contextually suitable replies. GOAT-SLM significantly outperforms all other models on both tasks.

Table 8: Results (%) on the Paralinguistic task.

Model	**ESD-zh**	**Para_mix300-zh**	**Age-zh**
GLM-4-Voice	35.55	1.89	27.81
MiniCPM-o 2.6	44.03	2.08	34.56
Baichuan-Omni-1.5	13.55	1.80	12.24
SpeechGPT-2.0-preview	22.59	1.52	23.63
Freeze-Omni	20.72	1.85	13.68
Qwen2.5-Omni	44.83	2.19	42.51
Kimi-Audio	53.17	9.19	22.77
GOAT-SLM	45.31	40.91	72.13

We also evaluate the quality of audio responses and the models' ability to generate speech. Following the setup in TELEVAL, we use the ESD-zh dataset to assess three aspects: DNSMOS, CER, and Emotion. As shown in Table 9, while models perform similarly in terms of DNSMOS scores, GOAT-SLM consistently outperforms all other open-source models across the remaining speech-related metrics. These results highlight the advantages of its model architecture, training paradigm and GOAT-TTS's capability for high-quality data construction.

Table 9: Results of Spoken Response Generation Evaluation.

Model	**CER ↓**	**DNSMOS ↑**	**Emotion ↑**
GLM-4-Voice	6.58	3.46	31.66
MiniCPM-o 2.6	2.58	3.52	34.26
Baichuan-Omni-1.5	7.89	3.40	24.74
SpeechGPT-2.0-preview	17.27	2.46	27.48
Freeze-Omni	4.88	3.49	41.05
Qwen2.5-Omni	1.69	3.47	52.59
Kimi-Audio	3.84	3.38	45.48
GOAT-SLM	1.57	3.46	61.48

In addition, we evaluate dialectal speech generation using the Chitchat-dialect dataset. In the subjective evaluation, 10 native speakers are recruited for each dialect to determine whether the model-generated speech belong to the target dialect. Since other baseline models lack dialect follow capabilities, their samples are excluded from the subjective assessment. As shown in Table 10, despite underperformance in northeastern Mandarin, GOAT-SLM achieves superior performance (with consistency rates exceeding 90%) in the remaining four dialects. Through analysis of the evaluation results, we find that even though the text generated by the Write module does not exhibit dialect features, the generated speech can still effectively acquire dialect-specific cues through the representations from the Listen module. This phenomenon validates the effectiveness of our proposed dual-modality head architecture design: while maintaining independent optimization of the text generation module and the speech generation module, the implicit interaction through deep representations successfully achieves effective transmission of dialect characteristics.

Table 10: Subjective Evaluation Results for Dialectal Spoken Response Generation.

Model	**Dialectal Spoken Response (%) ↑**				
	cantonese	**henan_dialect**	**northeastern_mandarin**	**shanghainese**	**sichuanese**
GOAT-SLM	97.4	99.6	57.2	99.0	92.6

6 Conclusion

In this work, we presented GOAT-SLM, an end-to-end spoken language model designed to support natural and adaptive voice-based interaction. The model features a dual-modality head architecture that decouples linguistic modeling from speech generation, and a modular, progressive training strategy that integrates large-scale pretrained language and speech models. Through targeted training across semantic, paralinguistic, and speaker-specific dimensions, GOAT-SLM is equipped to interpret and respond to both linguistic content and nuanced vocal cues. Evaluation results on the TELEVAL benchmark show that GOAT-SLM achieves balanced performance across multiple tasks and outperforms existing open-source models in several aspects of paralinguistic and speaker-aware interaction. While GOAT-SLM demonstrates strong capabilities in perceiving non-linguistic speech signals, further research is needed to enhance its fine-grained paralinguistic reasoning and its adaptability to more diverse and dynamic interaction scenarios.

References

[1] S. Arora, K. Chang, C. Chien, Y. Peng, H. Wu, Y. Adi, E. Dupoux, H. Lee, K. Livescu, and S. Watanabe, "On the landscape of spoken language models: A comprehensive survey," *arXiv Preprint*, 2025.

[2] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," in *Proc. EMNLP*, 2023, pp. 15 757–15 773.

[3] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," *arXiv Preprint*, 2024.

[4] Open-Moss, "Speechgpt 2.0-preview," https://github.com/OpenMOSS/SpeechGPT-2.0-preview, 2025.

[5] Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng, "Llama-omni: Seamless speech interaction with large language models," in *Proc. ICLR*, 2025, pp. 57 607–57 624.

[6] X. Wang, Y. Li, C. Fu, Y. Shen, L. Xie, K. Li, X. Sun, and L. Ma, "Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM," *arXiv Preprint*, 2024.

[7] W. Chen, Z. Ma, R. Yan, Y. Liang, X. Li, R. Xu, Z. Niu, Y. Zhu, Y. Yang, Z. Liu, K. Yu, Y. Hu, J. Li, Y. Lu, S. Liu, and X. Chen, "Slam-omni: Timbre-controllable voice interaction system with single-stage training," *arXiv Preprint*, 2024.

[8] A. Zeng, Z. Du, M. Liu, K. Wang, S. Jiang, L. Zhao, Y. Dong, and J. Tang, "Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot," *arXiv Preprint*, 2024.

[9] MiniCPM-o Team, OpenBMB, "Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone," https://github.com/OpenBMB/MiniCPM-o, 2025.

[10] Baichuan Inc., "Baichuan-omni-1.5 technical report," *arXiv Preprint*, 2025.

[11] W. Yu, S. Wang, X. Yang, X. Chen, X. Tian, J. Zhang, G. Sun, L. Lu, Y. Wang, and C. Zhang, "Salmonn-omni: A codec-free LLM for full-duplex speech understanding and generation," *arXiv Preprint*, 2024.

[12] Q. Chen, Y. Chen, Y. Chen, M. Chen, Y. Chen, C. Deng, Z. Du, R. Gao, C. Gao, Z. Gao, Y. Li, X. Lv, J. Liu, H. Luo, B. Ma, C. Ni, X. Shi, J. Tang, H. Wang, H. Wang, W. Wang, Y. Wang, Y. Xu, F. Yu, Z. Yan, Y. Yang, B. Yang, X. Yang, G. Yang, T. Zhao, Q. Zhang, S. Zhang, N. Zhao, P. Zhang, C. Zhang, and J. Zhou, "Minmo: A multimodal large language model for seamless voice interaction," *arXiv Preprint*, 2025.

[13] Qwen Team, "Qwen2.5-omni technical report," *arXiv Preprint*, 2025.

[14] Step-Audio Team, "Step-audio: Unified understanding and generation in intelligent speech interaction," *arXiv Preprint*, 2025.

[15] ——, “Step-Audio-AQAA: a fully end-to-end expressive large audio language model,” *arXiv Preprint*, 2025.

[16] Q. Fang, Y. Zhou, S. Guo, S. Zhang, and Y. Feng, “Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis,” *arXiv Preprint*, 2025.

[17] H. Shao, H. Gao, Y. Shen, J. Chen, L. Li, Z. Long, B. Tong, K. Li, and X. Sun, “Deeptalk: Towards seamless and smart speech interaction with adaptive modality-specific moe,” *arXiv Preprint*, 2025.

[18] Q. Wang, Z. Li, H. Lv, H. Chen, Y. Song, J. Kang, J. Lian, J. Li, Y. Li, Z. He, and X. Li, “BoSS: Beyond-semantic speech,” *arXiv Preprint*, 2025.

[19] Y. Shi, Y. Shu, S. Dong, G. Liu, J. Sesay, J. Li, and Z. Hu, “Voila: Voice-language foundation models for real-time autonomous interaction and voice roleplay,” *arXiv Preprint*, 2025.

[20] K. Team, “Kimi-audio technical report,” *arXiv Preprint*, 2025.

[21] Y. Song, H. Chen, J. Lian, Y. Zhang, G. Xia, Z. Li, G. Zhao, J. Kang, Y. Li, and J. Li, “GOAT-TTS: llm-based text-to-speech generation optimized via A dual-branch architecture,” *arXiv Preprint*, 2025.

[22] Z. Li, H. Chen, Y. Zhang, J. Zhou, X. Wang, H. Lv, J. Kang, J. Li, Y. Li, and X. Li, “TELEVAL: A dynamic benchmark designed for spoken language models in chinese interactive scenarios,” *arXiv Preprint*, 2025.

[23] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, “Qwen2-audio technical report,” *arXiv Preprint*, 2024.

[24] T. Li, J. Liu, T. Zhang, Y. Fang, D. Pan, M. Wang, Z. Liang, Z. Li, M. Lin, G. Dong, J. Xu, H. Sun, Z. Zhou, and W. Chen, “Baichuan-Audio: A unified framework for end-to-end speech interaction,” *arXiv Preprint*, 2025.

[25] Z. Wang, X. Liu, S. Liu, Y. Yao, Y. Huang, Z. He, X. Li, Y. Li, Z. Che, Z. Zhang, Y. Wang, X. Wang, L. Pu, H. Xu, R. Fang, Y. Zhao, J. Zhang, X. Huang, Z. Lu, J. Peng, W. Zheng, S. Wang, B. Yang, X. he, Z. Jiang, Q. Xie, Y. Zhang, Z. Li, L. Shi, W. Fu, Y. Zhang, Z. Huang, S. Xiong, Y. Zhang, C. Wang, and S. Song, “Telechat technical report,” *arXiv Preprint*, 2024.

[26] Y. Fathullah, C. Wu, E. Lakomkin, K. Li, J. Jia, Y. Shangguan, J. Mahadeokar, O. Kalinli, C. Fuegen, and M. Seltzer, “Audiochatllama: Towards general-purpose speech abilities for llms,” in *Proc. NAACL*, 2024, pp. 5522–5532.

[27] C. Wang, M. Liao, Z. Huang, J. Lu, J. Wu, Y. Liu, C. Zong, and J. Zhang, “BLSP: bootstrapping language-speech pre-training via behavior alignment of continuation writing,” *arXiv Preprint*, 2023.

[28] K. Deng, G. Sun, and P. C. Woodland, “Wav2prompt: End-to-end speech prompt learning and task-based fine-tuning for text-based llms,” in *Proc. NAACL*, 2025, pp. 6940–6956.

[29] K.-H. Lu, Z. Chen, S.-W. Fu, C.-H. H. Yang, J. Balam, B. Ginsburg, Y.-C. F. Wang, and H. yi Lee, “DeSTA2: Developing instruction-following speech language model without speech instruction-tuning data,” *arXiv Preprint*, 2025.