GOAT-SLM: 具有副语言和说话人特征感知的口语语言模型

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

近年来，端到端的口语语言模型（SLM）的进步显著提升了人工智能系统参与自然口语交流的能力。然而，大多数现有模型仅将语音视为语言内容的载体，往往忽视了嵌入在人类语音中的丰富副语言和说话者特征线索，如方言、年龄、情感和非语言声响。在这项工作中，我们介绍了 GOAT-SLM，一种具有副语言和说话者特征意识的新型口语语言模型，旨在将口语语言建模扩展到文本语义之外。GOAT-SLM 采用双模态头部架构，将语言建模与声学实现解耦，实现稳健的语言理解，同时支持富有表现力和自适应的语音生成。为提升模型效率和多功能性，我们提出了一种模块化的分阶段训练策略，利用大规模语音文本语料库逐步对齐语言、副语言和说话者特征信息。在 TELEVAL 的多维度评估基准测试上，实验结果表明 GOAT-SLM 在语义和非语义任务中获得了均衡的性能，并在处理情感、方言变化和年龄敏感交互方面优于现有的开源模型。这项工作强调了超越语言内容建模的重要性，并推进了更自然、自适应和具有社会意识的口语语言系统的发展。

1 介绍

近年来，端到端的口语语言模型（SLMs） [?] 取得了快速进展 [? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?] 。然而，大多数现有模型仍然主要将语音视为语言内容的载体，往往忽视了嵌入在音频信号中的丰富的副语言和说话者特征线索。这些线索——包括方言、年龄、情感和非语音声（NSV）信号如咳嗽——对于实现更自然、适应性更强和更具同理心的人机交互至关重要 [?] 。为弥补这一差距，我们提出了一种能够感知和响应此类非语言声信息的口语语言模型，通过提升对副语言和说话者特征的感知，增强交互质量。

在模型构建方面，SLM 通常遵循两种主要的范式。第一种是语音本地方法（例如，Moshi [?] ，SpeechGPT [?] ，SpeechGPT2 [?] ，Viola [?] ，Baichuan-Omni-1.5 [?] ），它直接将大规模语言模型（LLM）训练范式应用于大规模标记化语音语料库。第二种是模态对齐方法（例如，Freeze-Omni [?] ，Salmonn-Omni [?] ，MinMo [?] ，Kimi-Audio [?] ），它利用强大的 LLM 核心，通过跨模态对齐整合语音输入/输出模块。在这项工作中，我们采用模态对齐框架作为我们的基础，因为它具有计算效率并可以重用预训练的语音和文本模型。

---

† Equal contribution.

* Corresponding author. Email: xuelong_li@ieee.org

Technical report.

我们介绍了 GOAT-SLM，这是一种围绕四项核心原则设计的新型口语模型：生成导向的双模态设计、模块化训练的协调、对副语言和说话人特征的觉察以及文本智能的保留。如图 1 所示，GOAT-SLM 采用双模态头结构，其中预训练语言模型（LLM）的共享底层充当语义核心。这个核心分支为两个特定模态的头部：一个用于文本生成，另一个用于语音符号生成。为了促进参数共享和知识转移，语音头从文本头初始化。这种分叉结构带来了几个关键优势。首先，通过将语言模型与声学实现解耦，它保留了 LLM 的核心推理和理解能力。其次，通过利用丰富的语义表示，它可以实现高保真和富有表现力的语音合成。第三，该架构支持灵活的任务特定训练。例如，基于大规模 ASR 或 TTS 数据集进行训练可以产生强大的 TTS 性能。或者，通过在配对的语音到语音问答数据上进行训练，可以实现口语问答。事实上，我们扩展了这一框架以实现 GOAT-TTS [**?**]，用于生成合成的语音到语音对话数据来训练 GOAT-SLM。

为了训练，我们引入了一个由三个阶段组成的协调模块化策略：

- 指令微调：我们在用户指令中注入方言、年龄和情感等属性，使得 LLM 能够识别和响应细粒度的声学提示。这使得 LLM 具备生成更加同理心和上下文适宜的回应的能力。
- 模态对齐训练：使用重复和延续策略，我们在固定大语言模型（LLM）骨干网络的同时对齐语音和文本模态。这使得语音-文本的基础对齐成为可能，而不会降低 LLM 的预训练能力。
- 高保真语音生成优化：我们优化语音头以增强自然性、可理解性和表现力。在所有阶段中，保持 LLM 的文本智能，确保强大的综合推理和指令执行能力。

为了评估模型感知和响应副语言和说话者特征信号的能力，我们在一个综合对话评估框架 TELEVAL [**?**] 上进行了评估。评估结果显示，GOAT-SLM 在多个维度上表现出强大的性能，包括一般知识问答、声学鲁棒性、共情对话、方言互动以及年龄敏感的响应生成。在我们的演示网站 [1] 上可获得推理的对话样本。

2 相关工作

最近在大型语言模型（LLMs）方面的进展已显著影响了口语语言模型（SLMs）的发展，这些模型旨在以端到端的方式处理口语对话。虽然早期的工作主要集中于桥接语言和语音模态以实现语义理解，但最近的研究强调在将音频特定能力融入对话系统的同时需要保留核心语言智能。在本节中，我们回顾了这两条研究线的相关工作，并概述了我们的方法如何基于并扩展这些工作。

2.1 在口语模型中保持语言智能

将大型语言模型（LLMs）整合到语音语言模型（SLMs）中利用了从大规模文本语料库中获得的强大的语言理解和生成能力。然而，将 LLMs 适应语音任务时，通常存在灾难性遗忘的风险，即在对多模态数据进行微调后，模型的语言能力会下降。解决这一挑战已成为最近研究中的一个核心问题。

一种普遍的策略是课程式学习，其中 SLM 在大规模多模态数据集上进行持续的预训练，然后通过轻量级的指令微调进行微调。这种分阶段训练范式使模型能够获得音频理解能力，同时减轻其语言智能的退化。例如，SpeechGPT [**?**] 和 Moshi [**?**] 都采用了这种方法，表明持续的多模态预训练后进行有针对性的指令微调能够有效保持模型的语言能力。

另一种有效的策略是冻结 LLM 的核心参数，并通过辅助机制引入一小部分可学习的参数，以最大限度地降低干扰语言主干的风险。例如，MinMo [**?**] 在语音到文本对齐训练中采用基于 LoRA 的微调以保持 LLM 核心的稳定性。Freeze-Omni [**?**] 应用前缀调整来适应模态变化。DeepTalk [**?**] 在专家混合（MoE）架构中引入了音频特定的专家模块，以处理语音特定的信息。

我们的工作通过一个多分支建模框架实现了第二种策略，该框架用于联合文本和语音生成。具体来说，我们引入了一个并行的语音生成分支，使模型可以在不改变核心 LLM 架构的情况下同时处理这两种模态。值得注意的是，Kimi-Audio [**?**] 拥有类似的架构，并在我们的 GOAT-TTS 模型 [**?**] 大约同时发布。然而，不同于 Kimi-Audio 使用随机初始化的语音分支，

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

我们的方法使用预训练 LLM 的上层 Transformer 层初始化语音分支。该设计实现了更紧密的参数共享，并更好地保留了 LLM 的语言能力，同时扩展了其在语音感知对话生成方面的能力。

2.2 在大型音频语言模型中建模音频特定智能

尽管对保持语言智能给予了很多关注，SLM 设计的另一个重要方面在于建模音频特定的智能——解释和响应承载情境和情感信息的非语言声音线索的能力。最近的大型音频语言模型在这一方向上取得了显著进展，展示了识别各种基于音频的属性的能力，例如情感、语言/方言、说话者性别和说话者年龄 [? ? ? ?]。然而，这些努力主要集中在识别或描述用户的语音输入，缺乏基于检测到的信号自主调整响应的能力。因此，大多数模型仍然是反应性的，并依赖于提示，这限制了它们在更自然、情境感知的口语互动中的能力。

较小的一些研究开始探索响应式智能——即模型动态反应用户语音中副语言和与说话者相关的提示的能力。例如，最近的模型 [? ? ? ?] 引入了情感感知语音生成，允许根据用户指定的情感状态改变输出风格。然而，此类方法通常依赖于显式的控制信号，而不是自主的解释和上下文感知的响应。

我们的方法通过实现基于副语言和说话者特征意识的主动解释和自适应生成，在这一方向上取得了进展。该模型可以检测诸如方言特征、情感状态和非言语发声等线索，并能自主调整其响应，而无需明确指示。例如，当用户使用区域变体方言讲话时，模型可能以相匹配的方言进行回复，或者当检测到咳嗽或叹息等信号时，表现出同情心。

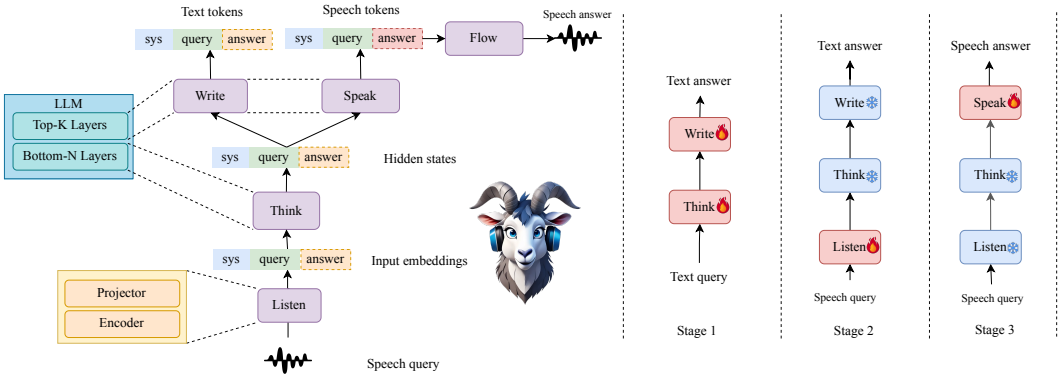

Figure 1: 框架与分阶段训练过程。阶段 1：指令微调。阶段 2：语音-文本对齐训练。阶段 3：高保真表达性语音生成训练。

3 架构

我们的模型架构，如图 1 所示，由五个功能模块组成——听、思考、写作、说话和流匹配——旨在支持统一的语音和语言理解与生成。这些模块协同工作，能够实现与副语言和说话者特征意识的无缝口语对话交互。

Listen 模块包含一个语音编码器和一个投影器，它们共同将输入语音转换为与 LLM 嵌入空间对齐的潜在表示。这使得语音输入能够在与文本相同的标记空间中进行处理，从而促进跨模态语义整合。

Think 模块由 LLM 的底部 N 层组成，作为共享的语义推理核心。Write 模块利用 LLM 的顶部 K 层来生成文本响应，形成标准文本交互的 Think–Write 路径。Speak 模块重复使用相同的顶部 K 层，但调整输出层以预测语音标记，从而实现语音生成的 Think–Speak 路径。

流匹配模块充当语音解码器，将预测的语音标记序列转换为声学特征（例如，声谱图），这些声学特征随后由声谱图到波形的声码器合成为波形。

在我们的实现中，我们采用 Whisper-small [1] 作为语音编码器，后接一个两层卷积神经网络和一个两层 Transformer 作为 Listen 模块中的投影器。对于语言骨干网络，我们使用 TeleChat2-

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

7B [**?**]，其中底部 15 层构成 Think 模块，顶部 15 层则被 Write 和 Speak 模块共享，分别用于生成文本和语音标记。

我们采用一个三阶段的训练策略，逐步让 GOAT-SLM 具备稳健的语音-文本对齐能力、副语言和说话者特征意识以及高保真表现力语音生成能力，如图 1 所示。GOAT-SLM 所建模的副语言和说话者特征信息如表 1 所示。

Table 1: GOAT-SLM 中的副语言和说话者特征信息。

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

我们仔细配置了包含文本查询（TQ）、语音查询（SQ）、文本应答（TR）和语音应答（SR）的不同组合的训练样本，具体细节见数据配置表 2 。每个阶段的超参数设置和训练成本详见表 3 。这种训练数据的模块化设计允许每个阶段针对特定能力，同时保持一致的整体学习目标。

Table 2: 训练的数据配置。

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

Table 3: 所有训练阶段的超参数。

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

为了使模型能够理解用户指令，同时感知并适应副语言和说话者特定属性，我们对明确标注有语音特征的多轮对话数据进行监督微调（SFT）。

训练数据中的每个用户查询由一个自然指令和一个描述性提示组成，该提示传达了年龄、方言、情感、非语音声事件（例如叹气、笑声）、语速和音量等声乐属性。这种设置为模型提供了关于说话者特征的明确线索，同时保留了对话意图的多样性，包括基于知识的问答、角色扮演、任务计划和开放式对话。对话以多轮格式结构化，以反映真实世界的互动动态。

为了增强模型对这些语音线索的敏感性，我们采用了一种对比数据构建策略：对于选定的指令，配对多种版本并附有不同的语音属性描述，鼓励模型产生可区分的响应。这种正负对比加强了语音特征与输出行为之间的一致性。

我们探索了多种提示格式来整合声乐属性，包括自然语言描述和基于关键词的标签，并使用特殊的标记来支持多属性组合，同时保持大型语言模型的流畅性。

模型生成的响应会经过进一步筛选和人工修正，以确保其自然性、连贯性和语境意识，同时仔细注意声音属性提示的一致整合。

该阶段建立了显式的说话者相关属性与自适应文本响应之间的对齐，为后续阶段的副语言感知口语建模提供了基础。

为了在语言和副语言维度上对齐语音和文本，我们采用了一种自蒸馏方法，在这种方法中，通过使用结构化转录提示 LLM 生成训练目标，类似于之前的工作 [**? ? ? ?**] 。我们使用了一个两阶段的渐进策略，以确保在这一阶段的稳定性和有效性。

阶段 2-1：语言对齐。我们首先在包含实际语音查询的大规模 ASR 数据集上进行训练。使用“重复和继续”提示方案生成相应的文本响应，其中 ASR 转录作为背景提供（见图 2 .a）。这

(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: 用于生成在构建语音文本数据对期间的文本响应的提示。文本、非语音、情感、语言和年龄标签分别表示与查询音频相关的转录、非语音发声、情感状态、语言或方言以及说话者的年龄。

一阶段侧重于在没有显式副语言信号的情况下进行稳健的语言对齐。思考-写作模块的参数和聆听模块的编码器被冻结，只有聆听模块中的投影器会被更新。

阶段 2-2: 语言学、超语言学和说话者特征对齐。在此阶段，语音查询由真实和合成话语的混合组成，这些话语根据多轮对话输入反映出各种超语言学和说话者特征。超语言学和说话者特征标签或通过人工标注，或通过自动分类器获得。对于每个训练样本，LLM 会被提供转录文稿和对应的属性描述 (图 2 .b) 以生成目标响应。在这一阶段，“听”模块的所有组件都经过微调，以实现与语言内容和语音属性的对齐。

3.1 阶段 3：高保真表达性语音生成训练

最终的训练阶段旨在赋予模型生成高质量、有表现力的语音响应的能力，这些响应能够反映丰富的副语言和特定说话者的特征。这是通过一个两阶段的程序实现的，阶段 3-1 和阶段 3-2，其公式与 GOAT-TTS 框架相似 [**?**] 。

阶段 3-1：冷启动语音生成训练。我们使用大规模真实数据集的数据驱动联合优化策略来启动 Speak 模块。对于普通话和英语，训练输入结构为 <Text Query, Text Response, Speech Response> 三元组，而方言数据则采用 <Speech Query, Text Response, Speech Response> 三元组（使用语音查询而非文本）。这一统一的范式有助于将语言内容和潜在属性（如情感）映射到语音输出，同时支持方言跟随。为了进一步增强长形式生成的鲁棒性，同一说话者的语音片段被随机串联形成最长可达 60 秒的样本。

阶段 3-2：属性感知的细化。这一子阶段增强了模型在风格适应性语音生成方面的能力。我们首先通过收集专业配音演员的录音构建多维语音提示集，涵盖四个核心情感（快乐、舒适、惊讶、中性）和三种目标听众群体（儿童、成人、老年人）。同时，从方言语料库中提取自然对话片段，并使用 GOAT-TTS 的流动模块将其转换为目标语音风格，以确保音色的一致性。

接着，我们构建训练三元组，方法是将来自阶段 2-2 的过滤样本——根据情感、年龄和语言分类——与收集的语音提示配对。使用 GOAT-TTS，将文本响应转换为语音目标以形成新的 <SQ, TR, SR> 样本。在这些数据上进行微调使得模型能够在不同的对话环境中生成具有情感表现、年龄感知和方言敏感的语音输出。

训练配置和优化。在此阶段，Listen 模块和 Think 模块的参数被冻结，训练集中于 Speak 模块。通过保留 Think 模块，模型保持了强大的文本理解能力，使 Speak 模块能够处理由 Write 模块生成的噪声或非标准文本输入，例如表情符号、换行符或特殊标点符号。

为了支持时间上连贯的语音合成，我们引入了一种多令牌预测（MTP）机制。在每个解码步骤中，将当前输出的令牌嵌入与下一步的输入特征相结合，形成一个融合的潜在表示，以指

导后续令牌的生成。这一机制结合双模态头架构，使得 Speak 模块能够以低延迟的方式进行流式语音生成，并在 Write 模块之后延迟一步。还采用了缓存管理策略，以在多轮对话中保持效率和稳定性，而无需迭代微调。

值得注意的是，我们在训练过程中采用了一种基于置信度的自动梯度掩蔽策略：高置信度的语音标记接收完整的梯度更新，而低置信度的标记则被掩蔽。这种选择性优化显著增强了发音的稳定性和整体语音的保真度——这一技术也在 GOAT-TTS 中得到了利用。

4 评价

为了系统地评估 GOAT-SLM 的能力，我们使用 TELEVAL [1] 对该模型进行评估，这是一套多维基准测评套件，设计用于测量两个关键领域：(1) 语义智能和 (2) 副语言和说话者特征感知交互。语义智能评估包括多语言问答、方言问答和多轮对话，目标是评估模型在口头交互中理解、推理和准确回应的能力。副语言和说话者特征感知评估则涵盖情感条件反应生成、方言适应、年龄感知交互和非言语语音反应，重点是评估模型感知声音特征并生成自适应语音反应的能力。

4.1 语义智能的结果

为了评估模型的语义智能，我们首先采用音频问答（AQA）方法来测试它们在中文和英文的一般知识问答中的表现。如表 4 所示，没有开源模型在所有数据集上始终表现出色。然而，MiniCPM-o 2.6 [**?**] 、Qwen2.5-Omni [**?**] 和 Kimi-Audio [**?**] 在特定子集上表现强劲。对于 GOAT-SLM，由于整合了副语言和说话人特征识别，其在一般 AQA 方面的能力略有下降，但整体仍保持在平均水平。

Table 4: 中文和英文中的常见 AQA 任务结果（%）。

Model	LlamaQA-en	LlamaQA-zh	TriviaQA-en	TriviaQA-zh	WebQ-en	WebQ-zh	ChineseSimpleQA-zh	ChineseQuiz-zh
GLM-4-Voice [**?**]	67.67	53.00	34.89	27.00	37.00	34.62	14.47	47.09
MiniCPM-o [**?**] 2.6	70.67	58.33	46.95	30.59	48.50	39.42	13.68	46.25
Baichuan-Omni-1.5 [**?**]	69.33	58.00	34.89	29.75	42.98	39.32	15.74	51.09
SpeechGPT-2.0-preview [**?**]	0.00	36.33	0.12	13.62	0.00	20.33	4.16	27.12
Freeze-Omni [**?**]	66.00	57.67	37.87	23.78	41.95	35.60	14.48	49.76
Qwen2.5-Omni [**?**]	69.67	58.67	43.13	29.03	44.32	35.19	13.42	56.30
Kimi-Audio [**?**]	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

我们进一步评估了模型在 AQA 任务中理解方言语音的能力，使用的音频输入涵盖五种中国方言：粤语、河南话、东北普通话、四川话和上海话。目的是评估当用户问题使用方言提出时，相较于在 ChineseQuiz-zh 数据集中使用标准普通话时，模型是否仍能提供准确的答案。如表 5 所示，Baichuan-Omni-1.5、Qwen2.5-Omni、Kimi-Audio 和 GOAT-SLM 对方言有一定程度的理解。然而，与 ChineseQuiz-zh 相比，它们的性能普遍有所下降。在这些方言中，东北普通话在大多数模型中表现最好，这可能是因为它与标准普通话的词汇差异相对较小。对于像粤语和上海话这样具有挑战性的方言，只有 Baichuan-Omni-1.5、Qwen2.5-Omni 和 GOAT-SLM 取得了相对较强的表现。值得注意的是，Baichuan-Omni-1.5 和 Qwen2.5-Omni 经过了比我们的训练语料大 4 倍的音频数据训练，这可能有助于它们在处理方言变化时的稳健性。

Table 5: 粤语、河南方言、东北官话、上海话、四川话的方言 AQA 任务的结果（%）。

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

对模型处理和保留多轮对话信息能力的评估使用了一个包含 150 个构建的多轮对话的数据集。为了防止在中间轮次中由于可能无法控制的模型响应导致的语义漂移，我们设计对话

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

时，让用户在早期轮次中提供信息，而模型只需确认输入。在最后一轮中才提出实际的问题。如表 6 所示，大多数模型表现出与基于文本的 LLM 对应模型相当的多轮推理能力。值得注意的是，GOAT-SLM 尽管没有经过任何多轮对话数据的训练，仍然表现出竞争力。这可能归因于其在当前训练模式下，输入嵌入的一致性较好。

Table 6: 多轮对话任务的结果（%）。

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

首先，我们评估模型在方言跟随任务中感知和响应副语言特征以及说话者特征的能力。与方言 AQA 任务相似，我们使用相同的五种中国方言进行实验。如表 7 所示，此前在方言 AQA 任务中表现卓越的开源模型（表 5 ）在此设定中表现明显较差。这表明，尽管模型可能理解方言内容，但通常无法在没有额外指令的情况下识别和自然地模仿用户的方言风格。GOAT-SLM 继续展现出与其在方言 AQA 任务中一致的强劲表现。这表明，该模型不仅能够理解方言输入并进行问答，还能够在开放域对话场景中生成适当的方言响应，而无需任何明确的指令。

Table 7: 方言跟踪任务的结果（% ）。

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

表格 8 展示了在三个副语言维度上的评估结果：情感、非语音声信号（NSV）和说话者年龄。在 ESD-zh 数据集中，用户的发言不包含任何明确的文本情感提示，使我们能够评估模型是否可以从语音语调中推断用户的情感状态并作出适当回应。Para_mix300-zh 测试集通过结合四种类型的 NSV 信号扩展了标准的 AQA 风格问题，旨在评估模型是否能够检测并响应此类提示。在 Age-zh 数据集中，查询通过合成的儿童或老年人语音进行发声，并且问题的内容经过调整，以适合相应的年龄段。

在评估这三个数据集时，我们不仅关注模型是否能够识别或分类副语言特征，更重要的是关注它们是否能够生成在上下文和社交上合适的回应。结果显示，当用户的情感或年龄在输入中有所体现时，大多数模型能够生成合理合适的回应。然而，所有开源模型在回复中都未能有效处理非语言的声信号。在这些模型中，Kimi-Audio 表现出一些检测这种信号的能力，但仍未能生成上下文适合的回应。在两个任务上，GOAT-SLM 明显优于其他所有模型。

Table 8: 副语言任务的结果（%）。

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

我们还评估了音频响应的质量以及模型生成语音的能力。按照 TELEVAL 中的设置，我们使用 ESD-zh 数据集来评估三个方面：DNSMOS、CER 和情感。如表 9 所示，在 DNSMOS 得分方面，各模型表现相似，而 GOAT-SLM 在其余的语音相关指标上始终优于所有其他开源模型。这些结果突出了其模型架构、训练范式以及 GOAT-TTS 在高质量数据构建能力上的优势。

Table 9: 口语响应生成评估的结果。

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

此外，我们使用 Chitchat-dialect 数据集评估方言语音生成。在主观评价中，为每种方言招募了 10 名母语者，以确定模型生成的语音是否属于目标方言。由于其他基线模型缺乏方言跟随能力，其样本被排除在主观评估之外。如表 **??** 所示，尽管在东北官话方面表现不佳，GOAT-SLM 在其余四种方言中表现优异（连贯性率超过 90%）。通过对评估结果的分析，我们发现尽管文本生成模块生成的文本不显示方言特征，生成的语音仍可以通过听模块的表示有效获取方言特定的提示。这一现象验证了我们提出的双模态头架构设计的有效性：在保持文本生成模块和语音生成模块独立优化的同时，通过深度表示的隐式交互成功实现了方言特征的有效传输。

在这项工作中，我们提出了 GOAT-SLM，一种端到端的口语模型，旨在支持自然和自适应的语音交互。该模型具有双模态头结构，可将语言建模与语音生成分离，并采用模块化、渐进式训练策略，集成了大规模预训练的语言和语音模型。通过在语义、语旁和说话者特定维度上的有针对性的训练，GOAT-SLM 能够解读和响应语言内容以及细微的语音线索。在 TELEVAL 基准上的评估结果显示，GOAT-SLM 在多个任务中表现均衡，并在语旁和说话者感知交互的若干方面优于现有的开源模型。尽管 GOAT-SLM 在感知非语言语音信号方面表现出色，但仍需进一步研究以增强其细粒度的语旁推理能力及其对更具多样性和动态性交互场景的适应能力。