arXiv:2603.09627v1 [eess.AS] 10 Mar 2026

SPEECH-OMNI-LITE：面向视觉语言模型的便携式语音接口 S

Dehua Tao [1] Xuan Luo [2 3] Daxin Tan [1] Kai Chen [4]
Lanqing Hong [1] Jing Li [2] Ruifeng Xu [3 5] Xiao Chen [1]

摘要

尽管大规模全模态模型已在多种模态上展现出令人印象深刻的能力，但其强大性能严重依赖海量多模态数据，并带来巨大的计算成本。本文提出SPEECH-OMNI-LITE，一种经济高效的框架，用于为预训练视觉语言（VL）骨干网络扩展语音理解与生成能力，同时完全保留骨干网络的视觉语言性能。具体而言，视觉语言骨干网络配备两个轻量、可训练的即插即用模块，即语音投影器和语音词元生成器，同时保持视觉语言骨干网络完全冻结。为缓解口语问答语料库的稀缺，本文提出一种低成本数据构建策略，从现有自动语音识别语音-文本对生成问题-文本答案-文本-语音（QTATS）数据，从而促进有效的语音生成训练。实验结果表明，即便仅使用数千小时的语音训练数据，SPEECH-OMNI-LITE也能取得优异的口语问答性能，可与在数百万小时语音数据上训练的全模态模型相媲美。此外，所学语音模块在不同视觉语言骨干网络之间展现出很强的可迁移性。

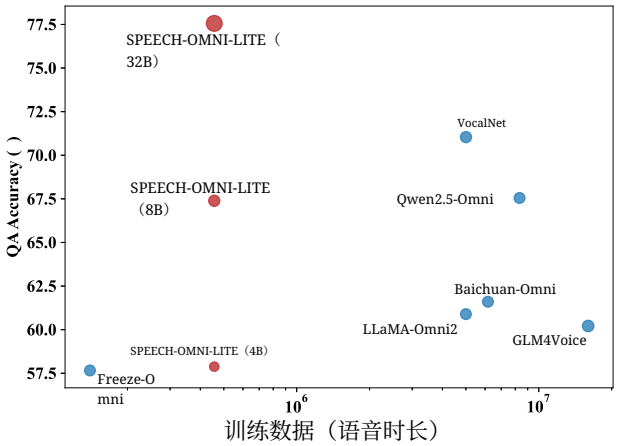


图 *1.* 不同模型的训练成本与问答准确率对比。语音时长通过汇总用于赋予模型语音理解和生成能力的所有语音数据计算得出。我们的方法以约十分之一的训练成本实现了有竞争力的性能。

1. 引言

在大规模模型时代，人们对能够在单一基础模型中统一文本、图像、语音和视频的多模态模型越来越感兴趣，这类模型被称为**全模态模型**。尽管全模态模型（OpenAI，2024年；Li 等，2025；QwenTeam，2025b）在各个模态上都展现出卓越的性能，但其成功通常依赖于大规模训练，这需要海量多模态数据集和大量计算资源。这给大多数研究团队带来了实际障碍，也引出了一个注重资源的问题：我们如何以经济高效的方式，将训练充分的基础模型扩展到新的模态，同时尽可能减少数据获取和计算开销？

扩展模型模态的一种常见策略是以基于文本的大语言模型（LLM）为基础，添加特定模态的组件，然后对整个模型进行微调，以实现跨模态对齐。这一范式已广泛应用于将语言模型扩展至视觉领域，催生了视觉–语言（VL）模型，例如 LLaVA（Liu 等人，2023）和 Qwen2-VL （Wang 等人，2024 年），这些模型将视觉编码器和轻量级适配器与强大的语言骨干网络集成，从而赋予模型视觉感知和推理能力。

除了视觉模态之外，**语音**也是人机交互的主要自然媒介，促使研究人员将其融入大语言模型，使这些模型具备“聆听”（语音理解）和“说话”（语音生成）的能力（Fang 等人，2025；Wang 等人，2025b）。现有

[1]华为莱布尼茨研究中心，香港特别行政区 [2]香港理工大学，中国香港特别行政区 [3]哈尔滨工业大学（深圳），中国深圳 [4]香港科技大学，中国香港特别行政区 [5]深圳环湾研究院，中国深圳。通信作者：Xiao Chen <chen.xiao2@huawei.com>。

预印本。*2026*年*3*月*11*日。

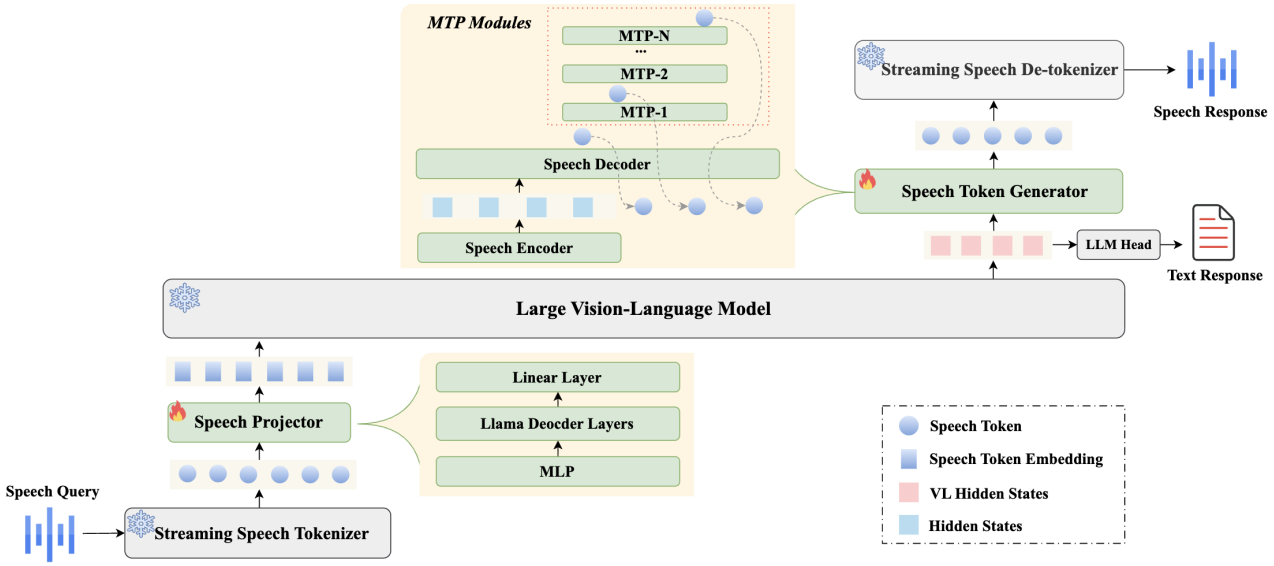


图2。 SPEECH-OMNI-LITE 的架构。它由一个预训练离散语音分词器、一个可训练的语音投影器、一个预训练的大型视觉-语言模型、一个可训练的 语音词元生成器以及一个预训练语音反分词器组成。

方法大致可分为两类。**（i）词元输入、词元输出**方法将语音离散化为类似文本词元的词元序列，并训练统一的语音-语言模型来接收和生成离散词元（小米，2025；Zeng 等人，2024）。这种方法为大语言模型提供了简洁的接口，但通常需要大规模语音-文本对齐数据和大量训练。**（ii）嵌入输入、词元输出**方法将语音编码为连续表示，并使用适配器将其映射到大语言模型的输入空间；随后，语音词元解码器根据大语言模型的隐藏状态生成语音词元（Chen 等人，2025b；QwenTeam，2025b）。尽管效果良好，这些方法通常至少需要对大语言模型进行部分微调以实现对齐，这可能会显著改变（甚至降低）骨干模型性能，并带来灾难性遗忘的风险（Luo 等人，2025）。此外，学习得到的语音适配器和生成器往往与特定的微调骨干网络紧密耦合，因此难以迁移到新的骨干网络，除非重新训练。

在这项工作中，我们提出了**SPEECH-OMNI-LITE**，这是一个经济高效的框架，可为预训练的视觉语言骨干网络扩展语音理解和生成能力，**同时保持视觉语言骨干网络完全冻结**。我们的方法（图2）首先使用预训练离散语音分词器将输入语音分词为离散单元。接着，轻量级的语音投影器将这些离散词元映射到视觉语言输入嵌入空间，使冻结的视觉语言骨干网络能够处理语音输入。在输出端，语音词元生成器将视觉-语言隐藏状态转换为离散语音词元，随后由预训练语音反分词器将其转换为波形。通过将语音专属学习限制在冻结的视觉语言骨干网络之上的轻量级可训练模块中，SPEECH-OMNI-LITE 保留了原有的视觉语言能力，同时实现了语音交互。

除了模型设计方面的挑战，数据稀缺也是有效训练语音生成的一大障碍。当前的语音到语音对话系统通常使用口语问答语料库进行训练，而大规模收集这类语料库成本高昂。为减轻对这类数据的依赖，SPEECH-OMNI-LITE 利用**问题–文本答案–文本–语音（QTATS）**数据，这些数据通过大语言模型驱动的**反向问题生成**，基于丰富的现有自动语音识别语料库构建而成。这一流程无需专门收集口语问答录音或进行语音合成。为实现基于 QTATS 的语音词元生成器训练，我们进一步引入一个辅助文本投影器，将文本输入映射至视觉语言骨干网络的输入空间，从而通过文本条件监督有效生成语音词元。

SPEECH-OMNI-LITE 的主要贡献有三点：

- **扩展语音模态而不发生灾难性遗忘：**我们提出了一种框架，在冻结所有骨干网络参数的同时，为预训练的视觉语言骨干网络赋予语音理解和生成能力，从而保留其原生的视觉语言能力。
- **轻量且可迁移的语音模块：**我们设计了紧凑且可训练的语音模块（语音投影器和语音词元生成器），以及一个预训练的

语音分词器和一个反分词器。这种模块化设计大幅降低了训练开销，并能在不同的视觉语言骨干网络之间无缝迁移。

- **一种用于语音生成训练的新型低成本数据构建策略：**据我们所知，我们首次通过反向生成问题，将自动语音识别语音-文本对转换为口语问答数据。结合文本条件训练流程，QTATS 摆脱了对昂贵的大规模口语问答数据集的依赖，并支持对语音词元生成器进行有效训练。

2. 相关工作

2.1. 迈向真正的全模态模型：统一支持文本、视觉和语音

全模态模型旨在将多种模态统一到单个基础模型中，但现有模型在模态覆盖范围上差异显著。有些“全模态”模型主要专注于视觉-语言集成，却不原生支持语音；另一些则仅为大语言模型扩展了语音交互能力。例如，Om- niVL（Wang 等人，2022年）是早期基础模型，旨在联合理解和生成图像-文本与视频-文本内容，但并不原生支持语音模态。相比之下，LLaMA-Omni（Fang 等人，2024年；2025）通过为大语言模型添加语音编码器和解码器，专注于实现无缝的语音到语音交互，但并未纳入视觉感知。尽管这些模型推动了各自领域内的统一处理，但它们尚未达到“完全全模态”的程度，因为它们无法在输入和输出两端同时支持文本、语音和视觉模态。

较新的系统通过在统一架构中同时支持语音、文本和视觉（图像与视频），进一步迈向完全全模态。Baichuan-Omni- 1.5（Li 等人，2025）提出了一种端到端多模态框架，同时支持语音理解和语音生成，以及视觉感知与推理。Emova（Chen 等人，2025a）通过在单一模型架构中整合语音、视觉和语言生成，探索统一的多模态交互。值得注意的是，Emova 引入了端到端多模态对齐，尤其关注语音输出中的情感与语调表现力，从而应对自然对话交互的挑战。同样，Qwen3-omni（Qwen 团队，2025b）将通义千问模型系列扩展至文本、语音、图像和视频等模态，实现了所有模态输入与输出的全面支持，且不会造成性能下降。总体而言，这些模型超越了“文本输入、文本输出”的多模态处理流程，旨在实现完全多模态交互。

在这一领域，SPEECH-OMNI-LITE 进一步推动全面的全模态能力发展：为强大的视觉语言骨干网络赋予语音理解与生成能力，实现统一的多模态推理与交互，同时保留原有的视觉语言性能。

2.2. 语音模态集成策略

语音模态集成方法大致可分为基于流水线的方法和端到端方法。

语音模态集成最简单的部署方式是级联流水线，由自动语音识别（自动语音识别）、大语言模型和文本转语音（文本转语音）组成（Huang 等人，2024年）。这种设计有两个根本局限。首先，它使用文本作为中间表示，因此大语言模型只能接收自动语音识别提取的语义内容，而会丢弃可能有用的副语言线索和细粒度声学信息（Ward，2006；Kidd 等人，2011；Zhao 等人，2025）。其次，这种设计延迟较高且会累积错误：自动语音识别中的错误会传播，并在下游生成过程中被放大（Liu 等人，2025）。

端到端方法分为两种范式。第一种范式使用离散词元表示语音，并将其直接纳入大语言模型词汇表，从而实现文本和语音序列的统一下一个词元预测（Zhang 等人，2023；Zeng 等人，2024年）。这一框架在概念上简单明了，但要实现高质量语音生成，需要大规模语音词元化、精心设计词汇表，并进行大量联合训练。

第二种范式采用模块化**Thinker-Talker**架构（Wang 等人，2025b；QwenTeam，2025a；Fang 等人，2025年；Wang 等人，2025a），将语音转文本理解与文本转语音生成解耦。在该框架中，思考器（通常为大语言模型）处理由语音编码器编码、并通过适配器对齐到文本空间的语音输入，以进行语义理解与推理。说话器模块随后根据思考器的输出生成离散语音表示，这些表示再由词元到语音解码器转换为波形（Du 等人，2024年；Ren 等人，2024年）。这种解耦使每个组件针对其特定任务得到优化，从而提升可扩展性与生成质量。

SPEECH-OMNI-LITE 遵循 Thinker-Talker 范式，通过专用的语音理解与生成模块扩展预训练视觉语言骨干网络。它借助轻量级适配器集成语音，而非重新训练整个基础模型，从而以极低的训练成本实现高效的语音模态扩展。

3. 方法

SPEECH-OMNI-LITE 是一种语音到语音对话模型，它通过语音输入扩展预训练的视觉语言骨干网络

和输出，同时完整保留骨干网络的能力。图2展示了该架构，其由预训练离散语音分词器、可训练的语音投影器、预训练的大型视觉-语言模型、可训练的语音词元生成器以及预训练语音反分词器组成。在整个训练过程中，语音分词器、视觉语言骨干网络和反分词器均完全冻结，仅对轻量级语音投影器和语音词元生成器进行优化。

SPEECH-OMNI-LITE 通过语音投影器将离散语音令牌投影到视觉语言骨干网络的输入嵌入空间，从而实现语音理解并支持语音条件文本生成。另一方面，它通过语音词元生成器将视觉语言骨干网络的输出隐藏状态转换为离散语音令牌，从而实现语音生成。在推理时，这些模块与语音分词器和反分词器协同工作，通过理解语音输入并合成与视觉语言骨干网络生成的文本相对应的语音响应，支持端到端语音到语音交互。以下小节将详细介绍模型组件、训练流程和数据构建策略。

3.1. 用于文本生成的语音输入建模

3.1.1. 流式离散语音分词器

为支持离散语音词元输入和低延迟响应，SPEECH-OMNI-LITE 采用流式离散语音分词器，将输入语音逐块转换为词元。该分词器的标记速率为 12.5 Hz，由 HuBERT LARGE 编码器（Hsu 等人，2021）、卷积下采样层和有限标量量化（FSQ）模块（Mentzer 等人，2023）组成。HuBERT LARGE 编码器在 45 万小时的语音数据上进行预训练。由于 HuBERT 以 50 Hz 的频率生成帧级表示，因此引入两个卷积下采样层，将帧率降至 12.5 Hz。随后，FSQ 模块将连续表示映射为离散语音令牌。

为实现流式训练和推理，多个组件经过重构，以便对音频块进行因果处理，包括：（i）基于卷积的模块（HuBERT 中的 7 层卷积特征提取器和两个下采样卷积层），以及（ii）Transformer层，其中采用了因果注意力和分块式位置嵌入。在训练过程中，FSQ 编码向量被输入一个 4 层卷积网络，并使用音素标签通过连接主义时间分类（CTC）损失对输出进行优化（Graves 等人，2006年）。该卷积堆栈充当仅用于训练分词器的辅助预测头；训练完成后将其丢弃，在提取离散语音令牌时不会使用。训练过程遵循以往的离散标记化方法（Tao 等人，2024年；Chen 等人，2025a；Zhang 等人，2026年）。

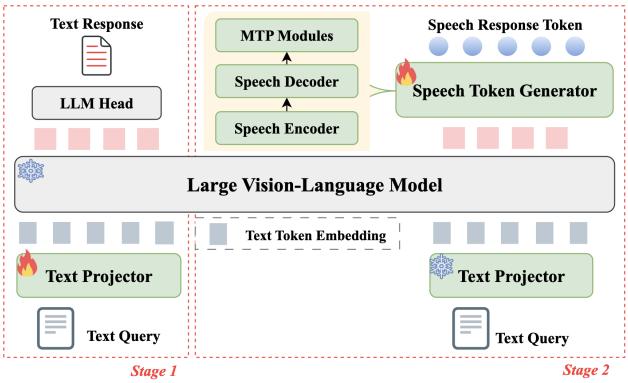


图3. 语音词元生成器的训练策略。第一阶段，在冻结视觉语言骨干网络的情况下训练文本投影器。第二阶段，在冻结其他模块的情况下训练语音词元生成器。两个阶段均使用 QTATS 数据。

架构细节见附录A。

3.1.2. 语音投影器

语音投影器将语音词元嵌入映射到视觉语言骨干网络模型的输入嵌入空间。如图2所示，它由一个多层感知机（MLP）、若干 LLaMA 解码器层和一个最终线性投影组成。多层感知机将语音词元嵌入维度提升至 Transformer 的隐藏维度，最终线性层则将隐藏状态映射为与视觉语言模型输入嵌入维度相匹配的表示。该投影器与视觉语言骨干网络连接，以支持流式语音输入。

语音投影器采用两阶段方案进行训练。首先，使用自动语音识别语音-文本对进行训练，使语音输入与文本输出对齐。其次，进一步使用语音问题-文本答案数据进行训练，以实现语音问答。在整个训练过程中，仅更新语音投影器参数，而离散语音分词器和视觉语言骨干网络则始终完全冻结。

3.2. 用于语音合成的语音输出建模

3.2.1. 语音词元生成器

为了从视觉语言骨干网络的输出隐藏状态中生成离散语音令牌，我们采用编码器-解码器架构实现了语音词元生成器，其设计遵循 VocalNet（Wang 等人，2025b）和 Freeze-Omni（Wang 等人，2025a）中的先前设计。此外，我们还引入多词元预测（Wang 等人，2025b）以加速解码。如图3所示，首先将视觉-语言隐藏状态输入线性投影层，然后由编码器进行处理。在编码器输出的条件下，解码器预测下一个语音词元，随后由若干

轻量级 MTP 头预测后续一小段词元。该解码器-MTP 流程以自回归方式重复，直至生成完整的词元序列（例如，输出序列结束词元或达到预设长度时）。

问题文本-答案文本-语音（QTATS）数据：语音词元生成器经过训练，可将视觉-语言隐藏状态映射为语音令牌。一种直接的监督策略是使用自动语音识别语料库，将语音作为输入，并将对应的离散语音令牌作为目标。然而，在我们的实验中，仅使用自动语音识别式条件输入训练的生成器在问答场景中的泛化能力较差：它能够可靠地重建自动语音识别语句中的词元，但在以问答过程中产生的视觉-语言隐藏状态为条件时却无法奏效。一种可能的解释是存在条件不匹配：由自动语音识别式条件输入产生的视觉-语言隐藏状态，与由问答式条件输入产生的隐藏状态有所不同，即使两者最终的文本内容相同或高度相似。与此同时，口语问答语料有限，通过文本转语音大规模合成口语问答的成本可能很高。为获得可扩展的监督数据，同时避免专门收集口语问答数据，我们利用现成的自动语音识别语音-文本对构建 QTATS。具体而言，对于每个自动语音识别语音-文本对（x^{sph}，y^{txt}），将转录文本 y^{txt}视为答案文本。我们提示一个大语言模型根据该答案生成相应的问题文本 。原始语音 x^{sph} 保留为答案语音，从而得到三元组（q^{txt}，a^{txt}，a^{sph}），用于在问答式条件输入下训练语音词元生成器。

辅助文本投影器：为了有效利用 QTATS，我们引入了一个辅助文本投影器，其架构与语音投影器相似，但处理的是文本输入。首先，使用视觉语言骨干网络的文本分词器对文本进行分词，然后通过骨干网络的输入嵌入矩阵将其映射为嵌入。这些嵌入被输入文本投影器，该投影器与冻结的视觉-语言骨干网络相接，从而无需口语提问即可实现问答式条件输入。

训练策略：图3 展示了训练过程，该过程分为两个不同阶段。首先，文本投影器在视觉语言骨干网络保持冻结的情况下进行训练，以支持处理 QTATS 问题文本，使骨干网络生成适用于问答式生成的隐藏状态。其次，连接语音词元生成器，并在 QTATS 上进行训练，将由此产生的视觉-语言隐藏状态映射为与答案语音对应的目标语音词元序列，同时保持视觉语言骨干网络和文本投影器冻结。训练完成后，不再需要文本投影器，可以将其移除。

表 *1*。自动语音识别数据集统计

语言	数据集	时长（小时）
中文	WenetSpeech	500
	MagicData-RAMC	128
	内部数据	1,500
英语	LibriSpeech	500
	GigaSpeech	500
	LibriHeavy	500
	CommonVoice	500
总计		4,128

3.2.2. 语音反分词器

语音去令牌化会根据语音词元生成器预测的离散语音令牌合成波形音频。去令牌器总体沿用了 F5-TTS（Chen 等人，2025c）的设计，但将其 DiT 模块替换为一种交叉注意力增强变体，称为 CA-DiT。具体而言，带噪声且经过掩码处理的声学表示会拼接在一起，并作为主要输入传入 CA-DiT 模块，而语音令牌则通过交叉注意力（作为键值对）充当条件信号。为缓解声学特征与词元序列之间的帧率不匹配（例如，梅尔特征的帧率为~94 Hz，而词元的帧率为 ~12.5 Hz），在进行条件化之前，会对词元序列进行时间上采样，使其与梅尔帧率更加匹配。架构细节见附录 B。

4. 实验设置

4.1. 数据集

在本研究中，我们使用约 4,000 小时的语音-文本配对自动语音识别数据（涵盖中文和英语）来训练语音分词器和语音投影器。这些语料库取自 6 个开源数据集（2 个中文数据集和 4 个英语数据集）；此外，我们还补充了 1,500 小时的内部中文数据，以确保最终语料库中的中文和英语数据量大致均衡，数据集级别的统计数据见表 1。训练语音分词器时，CTC（Graves 等人，2006）监督信号通过使用字素转音素模型将每条转录文本转换为音素序列而得到。对于中文，生成的音素直接用作 CTC 目标。对于英语，则将 BPE 应用于每个单词内的音素序列，并将词内 BPE 片段用作 CTC 目标。训练语音投影器时，视觉语言骨干网络的输出监督信号是与输入语音对应的转录文本。

为从自动语音识别语料库构建第 3.2.1 节所述的 QTATS 数据，我们使用通义千问3-30B-A3B-Instruct-2507[1] 来为每个转录文本可作为答案的语音-文本对生成一个问题。这一过程最终生成约

[1]https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507

145 万个 QTATS 问答对，用于训练辅助文本投影器和语音词元生成器。此外，为进一步提升 SPEECH-OMNI-LITE 的语音问答能力，我们使用了 VoiceAssistant-430K 数据集。该数据集是 VoiceAssistant-400K 的清理版本，移除了过长的响应，包含 43 万个语音查询-文本-响应对。

语音反分词器使用约 6k 小时的语音数据进行训练，其中包括 LibriTTS（Zen 等人，2019年）、Wenet-Speech4TTS（Ma 等人，2024年）以及内部数据。使用训练好的语音分词器，在分块推理设置下从音频中提取离散语音令牌，并将其作为条件输入用于反分词器训练。

4.2. 模型配置

流式离散语音分词器： HuBERT LARGE 用作前端语音编码器，后接两个卷积层，整体进行 $2\times$ 时间下采样，并使用一个量化级别为 $[8,8,8,8,8]$ 的 FSQ 模块。引入一个四层卷积音素解码器（步幅为 1）作为辅助头，将 FSQ 编码向量映射为适用于计算 CTC 损失的表示。语音块大小设为 640 毫秒，在流式训练期间，每个语音块都会关注前 30 个语音块。语音词元提取使用相同配置。除非另有说明，音素解码器仅用于分词器训练，并在词元提取时丢弃。语音分词器（不包括音素解码器）约含 337M 个参数，输出频率为 12.5 Hz 的词元。

语音投影器： 语音投影器由以下部分组成：(i) 一个具有 GELU 激活函数的 2 层多层感知机，可将每个 5 维 FSQ 代码向量映射为 2048 维表示；(ii) 6 个 Llama 风格解码器层，隐藏维度为 2048，包含 32 个注意力头，前馈网络维度为 8192；以及 (iii) 一个最终线性层，用于将解码器输出投影到视觉语言输入嵌入维度（即与骨干网络的输入嵌入大小相匹配）。语音投影器总共包含约 4.15 亿个参数，其中包括最终投影层的参数。

视觉语言骨干网络： 本研究使用 Qwen3-VL-8B-Instruct[2] 作为骨干网络。由于整个训练过程中骨干网络始终完全冻结，SPEECH-OMNI-LITE 本质上与骨干网络无关：其组件可以集成到各种预训练骨干网络中，而无需重新训练或修改骨干网络本身（见第5.3节），从而便于在不同模型系列中实现可扩展部署。

语音词元生成器： 语音词元生成器的配置遵循 VocalNet 确立的框架。它采用编码器-解码器架构，包含一个 8 层语音编码器和一个 4 层语音解码器。两个组件均由采用 32 个注意力头的 Llama 风格解码器块实例化。隐藏维度和前馈网络（FFN）维度分别设为 4096 和 16384。该模型包含五个多词元预测模块，每个模块由一个 Llama 风格解码器层和一个线性输出头组成。推理时，多词元预测模块的数量减少至两个；根据实验结果，这种配置性能最佳。

语音反分词器： CA-F5-TTS 基于 F5-TTS，采用包含 22 层、16 个注意力头且嵌入维度/前馈网络维度分别为 1024/2048 的 CA-DiT 主干网络。为缩小语音词元与声学帧之间的速率差异，语音词元序列按 6 倍进行上采样，标记速率达到 75 Hz，从而更接近梅尔频谱图的帧率（∼94 Hz）。在我们的实验中，推理时以分块方式合成音频，分块大小为 25 个词元，对应 2 秒音频。

4.3. 训练与评估设置

训练： SPEECH-OMNI-LITE 分两个独立阶段进行训练：（i）语音投影器训练；（ii）语音词元生成器训练。语音投影器训练分两个阶段进行优化。阶段 1 使用完整的自动语音识别语料库，使模型具备自动语音识别能力。阶段 2 使用 10% 的自动语音识别语料库以及完整的口语问答语料库，使模型适应语音问答任务。语音词元生成器训练分两步使用完整的 QTATS 语料库：首先训练辅助文本投影器，然后在冻结文本投影器的情况下训练语音词元生成器。在两个训练阶段中，学习率均设为 2×10^{-4}，采用余弦退火调度，预热比例为 0.03。由于这两个阶段彼此独立，语音投影器和语音词元生成器可以使用同一个冻结的视觉-语言骨干网络分别或并行训练，并在训练完成后分别连接到骨干网络的输入端和输出端。

语音理解评估（S→T）： 语音理解通过自动语音识别和语音转文本问答任务进行评估，其中输入为语音音频，输出为文本。自动语音识别在三个中文测试集上进行评估，包括 *AISHELL-1* 测试集、*WenetSpeechtest-net*、*Wenet-Speech test-meeting*，以及两个英语测试集，包括*LibriSpeech test-clean*和*LibriSpeech test-other*。中文自动语音识别性能以字错误率（CER）报告，英语自动语音识别性能以词错误率（WER）报告。问答评估在*LLaMA Questions*、*Web Ques-tions*、*TriviaQA*和*AlpacaEval*上进行，准确率以 GPT-5 判定的正确答案百分比报告。

语音生成评估（S→S）： 语音生成在语音到-的相同四个问答集上进行评估

[2]https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct

表 2. **自动语音识别**和**问答**任务的结果。AS-1 是 AISHELL-1 的简称；WenetS. 是 WenetSpeech 的简称；LibriS. 是 LibriSpeech 的简称。"/"表示模型不具备自动语音识别能力。"*"表示仅训练语音投影器的最后一个线性投影层。GLM-4-Voice 的结果取自原论文，因为我们未能复现其自动语音识别功能。**粗体**表示各子组中的最佳结果，下划线表示第二佳结果。

模型	自动语音识别（%字错误率↓；%词错误率↓）					口语问答（%准确率↑）							
	AS-1	WenetS.（测试集）		LibriS.（测试集）		LLaMA Questions		Web Questions		TriviaQA		AlpacaEval	
	test	净会议			其他	S->T	S->S	S->T	S->S	S->T	S->S	S->T	S->S
GLM-4-Voice (9B)	2.46	-	-	2.82	7.66	76.67	63.67	49.50	46.90	49.50	44.84	65.15	58.59
VocalNet (8B)			/			83.33	**78.60**	62.90	59.80	62.16	**57.30**	75.76	**72.22**
LLaMA-Omni2 (7B)			/			79.33	72.00	51.40	49.20	48.70	44.84	64.14	63.64
Freeze-Omni（7B）			/			78.67	58.67	54.10	29.80	51.15	32.36	46.73	28.79
百川-Omni-1.5（7B）	2.37	8.70	9.86	2.54	5.63	77.67	65.67	49.90	32.39	54.70	47.28	64.14	13.92
Qwen2.5-Omni-7B	**1.12**	**6.20**	**7.80**	**2.31**	**4.19**	79.67	77.33	61.60	**60.10**	57.20	56.30	71.72	67.17
SPEECH-OMNI-LITE（8B）	5.77	16.54	18.32	4.39	9.22	85.00	73.33	51.50	46.20	56.80	50.40	76.26	65.66
SPEECH-OMNI-LITE (4B)*	9.01	19.92	20.74	5.76	10.81	78.33	-	43.70	-	41.30	-	68.18	-
SPEECH-OMNI-LITE (32B)*	5.46	16.50	18.18	3.58	7.75	**86.00**	-	**66.60**	-	**71.80**	-	**85.86**	-

语音设置。评估两个方面：（i）模型生成文本与合成语音之间的内容一致性，以及（ii）感知音频质量。合成语音使用 Whisper-large-v3 进行转录。[3] 内容一致性通过（1）GPT-5 对转录回答的评分以及（2）模型生成文本与转录文本之间的词错误率（词错误率）来衡量。音频质量使用 UTMOS（Saeki 等人，2022）。

5. 实验结果

5.1. 语音上下文到文本生成的结果

表2 报告了 SPEECH-OMNI-LITE 与具有代表性的语音语言模型和全模态模型在英语和中文评测集上的自动语音识别及语音转文本问答结果。为在流式设置下进行公平比较，对于所有支持流式语音输入的模型，流式分块大小均固定为 640 毫秒，与 SPEECH-OMNI-LITE 的配置一致。经验观察表明，一些以往的语音语言模型和全模态模型主要面向问答，缺乏自动语音识别能力（即 VocalNet、LLaMA-Omni2 和 Freeze-Omni[4]），这可能限制其对语音理解任务的覆盖范围。由于用于训练语音分词器和语音投影器的自动语音识别数据量相对较少，SPEECH-OMNI-LITE 在自动语音识别基准上的表现不如使用大量更多语音数据训练的大规模全模态模型（即 Baichuan-Omni-1.5、Qwen2.5-Omni）。尽管如此，SPEECH-OMNI-LITE 仍取得了出色的问答性能，尤其是在*LLaMAQuestions* 和*AlpacaEval*上。总体而言，这些结果表明，仅使用一个通过有限自动语音识别和问答数据训练的可训练语音投影器，SPEECH-OMNI-LITE 便兼具自动语音识别功能与有竞争力的口语问答性能，可与使用数百万小时语音训练的大规模全模态模型相媲美。

5.2. 语音上下文到语音生成的结果

文本与生成语音之间的内容一致性以及声学质量的评估结果汇总于表2 和表 3。与具有代表性的流式语音大语言模型和全模态模型相比，SPEECH-OMNI-LITE 的词错误率和 UTMOS 得分处于中等水平，但其语音到语音问答性能具有竞争力。值得注意的是，这一性能水平是在远比现有系统经济的监督设置下实现的。语音词元生成器完全通过人工构造的问答式监督进行训练，这些监督源自易于获取的自动语音识别语料库；相比之下，以往工作依赖大量经过精心整理或合成的口语问答对。由于这种人工构造的监督缺乏真正的问答推理结构，因此，在用于语音生成的视觉-语言隐藏状态条件中，训练—推理不匹配不可避免。尽管存在这一局限，实证结果表明，所提出的策略仍然非常有效——即使没有口语问答语料库，也能呈现稳健的语音生成行为。这一结果证明，仅依靠自动语音识别资源构建简单、可扩展且低成本的监督流程，就能为 Talker 结构实现高质量的语音生成。

5.3. 语音投影器的可迁移性

为考察使用 Qwen3-VL-8B-Instruct 骨干网络训练的语音投影器能否有效迁移至其他规模的视觉语言骨干网络，我们选取 Qwen3-VL-4B-Instruct[5]和 Qwen3-VL-32B-Instruct[6] 作为替代骨干网络。将训练完成的语音投影器应用于这些骨干网络时，仅更新最后的线性投影

[3] https://huggingface.co/openai/whisper-large-v3 [4] 尽管论文报告了自动语音识别的结果，但根据作者在 https://github.com/VITA-MLLM/Freeze-Omni/issues/10 中的回复，发布的模型无法执行自动语音识别任务。

[5] https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct [6] https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct

表 3. 语音响应与文本响应之间的**内容一致性**（词错误率 ↓）和**声学性能**（UTMOS ↑）结果。**粗体** 表示各子组中的最佳结果，下划线表示次佳结果。

模型	LLaMA Questions		Web Questions		TriviaQA		AlpacaEval		平均值	
	词错误率	UTMOS	词错误率	UTMOS	词错误率	UTMOS	词错误率	UTMOS	词错误率	UTMOS
GLM-4-Voice (9B)	41.00	4.15	10.16	4.11	13.14	4.15	29.44	3.86	23.44	4.07
VocalNet(8B)	2.79	**4.47**	5.01	**4.49**	4.86	**4.47**	7.43	**4.49**	5.02	**4.48**
LLaMA-Omni2（7B）	3.77	4.44	5.75	4.43	6.02	4.44	**3.88**	4.43	4.86	4.44
Freeze-Omni（7B）	24.34	4.36	47.31	4.23	43.61	4.26	49.47	4.14	41.18	4.25
百川-Omni-1.5（7B）	31.64	4.31	45.58	4.32	29.51	4.35	86.53	4.30	48.32	4.32
Qwen2.5-Omni-7B	**0.99**	4.30	**1.68**	4.32	**1.47**	4.29	5.54	4.20	**2.42**	4.28
SPEECH-OMNI-LITE（8B）	12.49	4.12	22.55	3.96	15.04	4.08	23.77	3.73	18.46	3.97

层，使其与骨干网络的输入维度相匹配，其余所有组件均保持冻结。训练数据和配置沿用语音投影器第二阶段训练时所用的设置。相应的语音理解结果见表 2。随着骨干网络规模增大，性能持续提升，其中 SPEECH-OMNI-LITE（32B）的语音到文本问答结果最佳。这些发现表明，在冻结主干网络设置下，使用较小视觉语言骨干网络训练的语音投影器可以有效迁移至更大的骨干网络，为其赋予语音理解能力，同时大幅降低原本为适配不同规模骨干网络所需的训练成本。

5.4. 语音响应流程的延迟分析

SPEECH-OMNI-LITE 的响应延迟可分为四个阶段：(1) 流式语音分词器将输入语音查询转换为离散词元；(2) 视觉语言骨干网络生成隐藏状态；(3) 语音词元生成器生成响应语音词元；(4) 语音反分词器合成输出波形。生成 1 秒语音响应的延迟汇总见附录 4 中的表格 C。语音反分词器是端到端延迟的主要来源。其延迟是在以下配置下计算的：根据语音词元以分块方式合成音频，每块大小为 25 个词元（对应 2 秒音频）。该系统的一项关键特性是，语音分词器完全以流式模式运行：每个 640 毫秒的语音块耗时固定为 54.3 毫秒，因此无论输入语音时长如何，延迟都保持恒定。

6. 讨论

语音投影器的架构影响：实验结果表明，当视觉语言骨干网络和语音分词器都保持冻结时，语音投影器的架构起着至关重要的作用。附录5 中的D 总结了在语音投影器第一阶段训练中获得的自动语音识别性能。结果表明，表达能力更强的投影器架构始终能带来更好的性能，其中基于Transformer的投影器在SPEECH-OMNI-LITE框架下取得了最佳结果。还观察到，较简单的架构需要显著更大的学习率才能有效训练（例如，MLP和基于卷积神经网络的投影器需要 10^{-3} 或 10^{-2}），而基于Transformer的投影器则能以较小的学习率（例如，2×10^{-4}）稳定训练。

语音词元生成器中的编码器–解码器配置影响： 语音词元生成器的两个主要组件——语音编码器和语音解码器——对生成的语音词元质量具有不同影响。附录6中的表 E的比较结果揭示了两个关键观察结果。首先，在Transformer层总数固定的情况下，为语音编码器分配更多层，相较于将这些层分配给语音解码器，能够带来更好的性能。其次，更深的语音编码器在弥合视觉语言骨干网络隐藏状态与目标语音词元嵌入之间的表示差距方面更为有效。这些发现凸显了编码器容量在将视觉语言表征与语音词元空间对齐方面的重要性。

7. 结论

本文提出了 SPEECH-OMNI-LITE，这是一种高性价比框架，可为预训练的视觉语言骨干网络增添语音理解和语音生成能力，同时保留其原有的视觉与语言性能。为避免依赖成本高昂的口语问答语料库，本文利用易于获取的自动语音识别资源构建 QTATA 数据，用于监督语音词元生成器的训练，为大规模音频合成提供了成本显著更低的替代方案。实验结果表明，即使仅使用数千小时的语音数据进行训练，SPEECH-OMNI-LITE 的口语问答能力仍可媲美主流语音语言模型和使用数百万小时语音数据训练的全模态模型。此外，使用一个骨干网络训练的语音投影器

可以有效迁移到不同规模的骨干网络上，并且随着骨干网络容量增大，语音理解性能持续提升。

影响声明

本文提出了 SPEECH-OMNI-LITE，这是一种旨在推动多模态机器学习领域发展的框架，通过为视觉-语言（VL）模型提供经济高效且便携的语音接口来实现这一目标。除学术贡献外，我们的工作还可能产生以下更广泛的影响和社会后果：

1. **多模态研究的民主化**。通过采用轻量级、“即插即用”的模块，并保持大型 VL 骨干网络冻结，我们的方法大幅减少了训练全模态模型通常所需的计算资源和海量数据集。这有助于降低硬件预算有限的小型研究团队和机构的参与门槛。

2. **资源效率与可持续性**。所提出的 QTATS 数据构建策略仅使用传统模型所需语音数据的大约十分之一，即可实现有效训练。训练时间和数据依赖性的减少，意味着模型开发阶段的碳足迹更小、能耗更低。

3.**无障碍与人机交互**。语音是人类交互的主要自然媒介。通过提供一种便携且高效的方式，为现有模型添加语音能力，本工作有助于开发更易访问的 AI 界面，供依赖基于语音交流的视觉或运动障碍人士使用。

4. **伦理考量与风险缓解**。 虽然我们的框架增强了模型能力，但也继承了大规模生成模型所面临的伦理挑战，例如生成有偏见或有害内容的可能性。不过，由于我们的方法在无需重新训练整个模型的情况下保留了原骨干网络的性能，因此能够避免“灾难性遗忘”，并继续使用冻结的视觉-语言骨干网络中已有的安全对齐机制。

参考文献

Chen, K., Gou, Y., Huang, R., Liu, Z., Tan, D., Xu, J., Wang, C., Zhu, Y., Zeng, Y., Yang, K., 等。Emova：赋能语言模型以鲜活的情感进行视觉感知、听觉感知和言语表达。载于计算机视觉与模式识别会议论文集，第 5455–5466 页，2025a。

Chen, Q., Chen, Y., Chen, Y., Chen, M., Chen, Y., Deng, C., Du, Z., Gao, R., Gao, C., Gao, Z., 等。Minmo：一种支持无缝语音交互的多模态大语言模型，2025b。网址 https://arxiv.org/abs/2501.06282。

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., JianZhao, J., Yu, K., 和 Chen, X. F5-TTS：一种通过流匹配生成流畅且忠实语音的讲述者。在计算语言学协会第 *63* 届年会论文集（第 *1* 卷：长篇论文），第 6255–6271 页，2025c。

Du, Z., Wang, Y., Chen, Q., Shi, X., Lv, X., Zhao, T., Gao, Z., Yang, Y., Gao, C., Wang, H., 等。Cosyvoice 2：基于大型语言模型的可扩展流式语音合成，2024年。网址https://arxiv.org/abs/2412.10117。

Fang, Q., Guo, S., Zhou, Y., Ma, Z., Zhang, S., and Feng, Y. Llama-omni：使用大语言模型实现无缝语音交互。 *arXiv*预印本 *arXiv:2409.06666*，2024年。Fang, Q., Zhou, Y., Guo, S., Zhang, S., and Feng, Y. LLaMA-omni 2：基于大语言模型的实时语音聊天机器人，采用自回归流式语音合成。在 Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T.（编），第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文），第18617–18629页，奥地利维也纳，2025年7月。计算语言学协会。ISBN 979-8-89176-251-0。doi: 10.18653/v1/2025.acl-long.912。网址: https://aclanthology.org/2025.acl-long.912/。

Graves, A., Fern´andez, S., Gomez, F., and Schmidhuber, J. 连接主义时间分类：使用循环神经网络对未分段序列数据进行标注。在第*23*届国际机器学习会议论文集，第369–376页，2006年。Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. HuBERT：通过对隐藏单元进行掩码预测来进行自监督语音表征学习。*IEEE/ACM*音频、语音与语言处理汇刊，29:3451–3460，2021年。

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J. 等。AudioGPT：理解并生成语音、音乐、声音和说话头像。在*AAAI*人工智能会议论文集，第38卷，第23802–23804页，2024年。

Kidd, C., White, K. S., and Aslin, R. N. 幼儿利用言语不流畅现象预测说话者的指称意图。发展科学，14(4):925–934，2011年。

Langley, P. 撰写机器学习论文。在 Langley, P.（编），第*17*届国际机器学习会议论文集（*ICML 2000*），第1207–1216页，加利福尼亚州斯坦福，2000年。Morgan Kaufmann。

Li, Y., Liu, J., Zhang, T. 等。Baichuan-Omni-1.5 技术报告，2025。网址https://arxiv.org/abs/2501.15368。

Liu, H., Li, C., Wu, Q. 和 Lee, Y. J. 视觉指令微调。神经信息处理系统进展，36:34892–34916，2023。

Liu, Z., Duan, Y., Wang, M., Feng, P., Zhang, H., Xing, X., Shan, Y., Zhu, H., Dai, Y., Lu, C., 等。X-talk：论模块化语音到语音对话系统被低估的潜力。*arXiv* 预印本 *arXiv:2512.18706*，2025年。Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J. 和 Zhang, Y. 大语言模型在持续微调过程中的灾难性遗忘实证研究。*IEEE* 音频、语音与语言处理汇刊，33:3776–3786，2025年。doi：10.1109/TASLPRO.2025. 3606231。Ma, L., Guo, D., Song, K., Jiang, Y., Wang, S., Xue, L., Xu, W., Zhao, H., Zhang, B. 和 Xie, L. Wenetspeech4tts：用于大规模语音生成模型基准测试的12,800小时普通话文本转语音语料库。载于*Interspeech 2024*论文集，第1840–1844页，2024年。Mentzer, F., Minnen, D., Agustsson, E. 和 Tschannen, M. 有限标量量化：让VQ-VAE变得简单。*arXiv* 预印本 *arXiv:2309.15505*，2023年。OpenAI。GPT-4o系统卡，2024年。网址：https://arxiv.org/abs/2410.21276。QwenTeam。Qwen2.5-Omni技术报告，2025a。网址：https://arxiv.org/abs/2503.20215。QwenTeam。Qwen3-omni技术报告，2025b。网址：https://arxiv.org/abs/2509.17765。Ren, Y., Wang, T., Yi, J., Xu, L., Tao, J., Zhang, C. Y. 和 Zhou, J. 采用时不变编码的少令牌神经语音编解码器。载于*ICASSP 2024—2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第12737–12741页。IEEE，2024年。Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S. 和 Saruwatari, H. UTMOS：东京大学SARULab的VoiceMOS挑战赛2022系统。载于国际语音通信协会年度会议论文集（*INTERSPEECH*），第2022卷，第4521–4525页，2022年。

Tao, D., Tan, D., Yeung, Y. T., Chen, X. 和 Lee, T. Toneunit：一种用于声调语言语音合成的语音离散化方法，2024年。网址：https://arxiv.org/abs/2406.08989。

Wang, J., Chen, D., Wu, Z., Luo, C., Zhou, L., Zhao, Y., Xie, Y., Liu, C., Jiang, Y.-G. 和 Yuan, L. Omnivl：面向图像-语言与视频-语言任务的统一基础模型。神经信息处理系统进展，35:5696–5710，2022。

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., 等。Qwen2-vl：提升视觉-语言模型在任意分辨率下对世界的感知能力，2024年。URLhttps://arxiv.org/abs/2409.12191。Wang, X., Li, Y., Fu, C., Zhang, Y., Shen, Y., Xie, L., Li, K., Sun, X., 和 MA, L. Freeze-omni：一种采用冻结大语言模型的智能低延迟语音到语音对话模型。载于第四十二届国际机器学习大会，2025a。URL https://icml.cc/virtual/2025/poster/43854。Wang, Y., Liu, H., Cheng, Z., Wu, R., Gu, Q., Wang, Y., 和 Wang, Y. VocalNet：采用多令牌预测、实现更快且高质量生成的语音大语言模型。载于 Christodoulopoulos, C., Chakraborty, T., Rose, C., 和 Peng, V.（编），*2025*年自然语言处理实证方法会议论文集，第 19584–19601 页，中国苏州，2025年11月。计算语言学协会。ISBN 979-8-89176-332-6。doi: 10.18653/v1/2025.emnlp-main. 989。
U
R
L
h
t
t
ps://aclanthology.org/2025.emnlp-main.989/。Ward, N. 美式英语中的非词汇性交谈声音。语用学与认知，14(1):129–182，2006年。Xiaomi, L.-C. Mimo-audio：音频语言模型是少样本学习器，2025年。URLhttps://arxiv.org/abs/2512.23808。Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., 和 Wu, Y. Libritts：一个从 LibriSpeech 派生的文本转语音语料库。*Interspeech 2019*，2019年。Zeng, A., Du, Z., Liu, M., Wang, K., Jiang, S., Zhao, L., Dong, Y., 和 Tang, J. Glm-4-voice：迈向智能、类人端到端语音聊天机器人，2024年。URLhttps://arxiv.org/abs/2412.02612。Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., 和 Qiu, X. Speechgpt：赋予大语言模型内在的跨模态对话能力。载于计算语言学协会自然语言处理实证方法会议论文集：*EMNLP 2023*，第 15757–15773 页，2023年。

Zhang, H., Tan, D., Tao, D., Chen, X., Tan, H., Li, Y., Cao, Y., Wang, J., 和 Song, L. Dsa-tokenizer：基于流匹配的层级融合解耦语义声学分词，2026年。URL https://arxiv.org/abs/2601.09239。

Zhao, X., Xu, Z., Cheng, Q., Fei, Z., Jin, L., Wang, Y., Chen, H., Jiang, Y., Gao, Q., Chen, K., 等。Moss-speech：迈向无需文本引导的真正语音到语音模型。*arXiv* 预印本 *arXiv:2510.00499*，2025。

A. 流式离散语音分词器

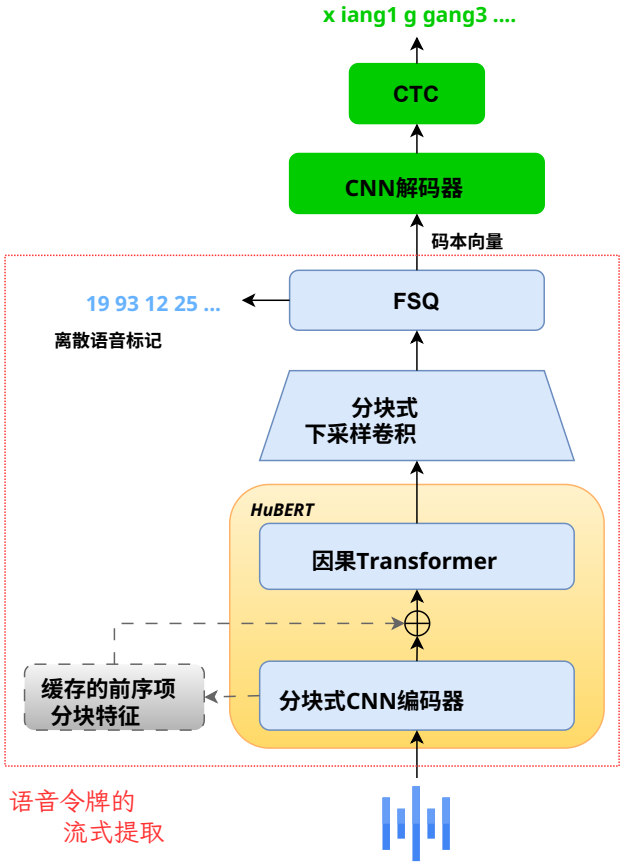


图 *4.* 流式离散语音分词器的架构。

语音分词器的训练方法可以视为一项任务：将预训练的 HuBERT LARGE 编码器微调为声调音素识别器。如图4所示，原始语音波形首先由 HuBERT LARGE 编码器进行处理，随后通过 FSQ 模块对连续表示进行量化，以获得码本向量。接着，这些量化嵌入被传递至基于 CNN 的音素解码器，该解码器预测音素标签。在训练过程中采用 CTC 损失，引导量化表示从语音信号中编码语音和词汇线索。这种设计确保生成的语音词元保留语音中包含的语义信息。

B. 语音反分词器

为了将离散语音令牌转换为连续的梅尔频谱图，我们设计了一种专用的语音反分词器，其框架类似于 F5-TTS，同时适配基于令牌的条件输入。该模型采用 CA-DiT 主干网络，如图 5 所示，并使用流匹配范式进行训练。

CA-DiT 与 F5-TTS 的 DiT 主干网络有两点关键区别：首先，它将文本输入替换为上采样后的离散语音令牌，作为主要的条件信号。其次，我们采用查询-键-值（QKV）条件注入机制：梅尔频谱图特征（作为查询）通过交叉注意力与离散语音令牌（作为键和值）交互，取代了 F5-TTS 将含噪梅尔频谱、干净梅尔频谱和文本特征拼接在一起的做法。这种解耦设计通过显式建模离散词元与连续梅尔帧之间的语义交互，增强了令牌-声学对齐。

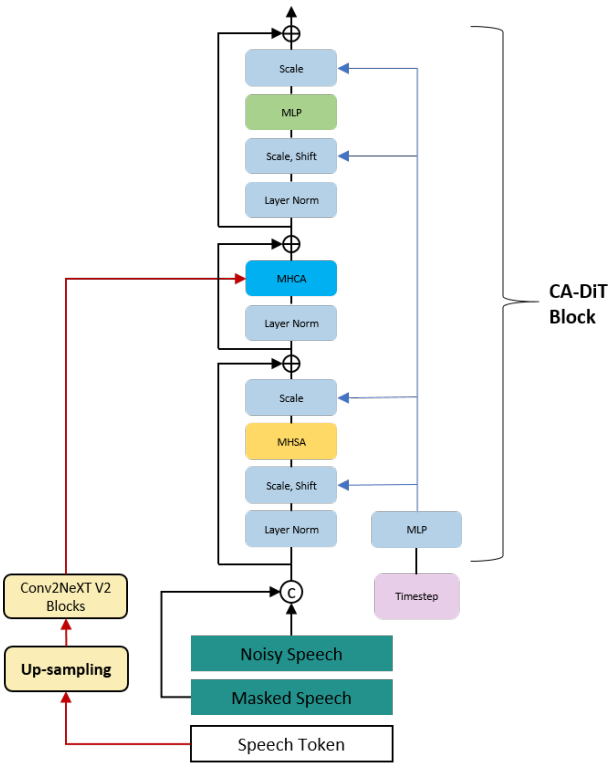


图*5*。 语音反分词器中 CA-DiT 模块的架构。

C. 端到端延迟分析

表4 报告了使用 SPEECH-OMNI-LITE（8B）生成 1 秒语音响应的端到端延迟。我们测量了响应流水线中各组件的延迟，包括流式语音分词器、VL 骨干网络、语音词元生成器和语音反分词器。

表4. 使用 Speech-Omni-Lite（8B）生成 1 秒语音响应的延迟分解（毫秒）。

语音分词器	VL	语音令牌生成器	语音反分词器	Sum
54.32	123.16	29.33	1139	1345.79

D. 语音投影器架构的详细分析

本附录提供了实验配置和完整的数值结果，以支持对语音投影器架构的比较。视觉语言骨干网络（Qwen2.5-3B[7]）和语音分词器保持冻结，仅使用自动语音识别数据优化语音投影器。评估的架构包括不同深度和学习率设置的多层感知机、卷积神经网络和基于Transformer的投影器。

表5 报告了实验中使用的模型结构、层数和学习率的确切组合，以及在 LibriSpeech (LS) test-clean 和 AISHELL2 (AS2) test 上的词错误率和字错误率结果。这些结果为分析不同架构选择下投影器的表达能力和训练稳定性如何变化提供了实证依据。

表5. 使用 Qwen2.5-3B 骨干网络时，不同语音投影器架构的自动语音识别性能。仅训练语音投影器，视觉语言骨干网络和语音分词器保持冻结。投影器的隐藏维度固定为 2048。

结构 语音投影器	层数 层	学习 Rate	词错误率（%）在 LS test-clean	字错误率(%)（针对） AS2测试
MLP	2	1e-2	19.4	28.6
	3	1e-2	11.6	19.5
	4	1e-2	失败	
	4	1e-3	9.2	20.3
	5	1e-3	12.3	20.5
CNN	4	1e-2	6.9	14.5
	4	1e-3	6.6	14.5
	4	2e-4	8.3	19.7
Transformer	2	1e-3	4.4	8.6
	2	2e-4	4.3	8.6
	4	1e-3	失败	
	4	2e-4	3.9	7.7
	6	2e-4	3.9	7.4

E. 语音词元生成器配置的详细分析

本附录报告了支持正文中层分配分析的实验配置和数值结果。训练时仅训练语音词元生成器，视觉语言骨干网络保持冻结。

表6 列出了针对 Qwen3-VL-4B 和 Qwen3-VL-8B 评估的精确编码器-解码器层配置，以及这些配置对应的词错误率。词错误率是根据模型生成的文本回复与使用 Whisper-large-v3 对生成语音进行转录所得的文本计算的。这些结果为分析编码器和解码器之间的层分配如何影响语音词元生成质量提供了实证依据。

表6. 语音词元生成器中语音编码器与语音解码器之间不同 Transformer 层分配方案的比较。视觉语言骨干网络被冻结，仅训练令牌生成器。每一行对应研究中评估的一种特定编码器–解码器深度配置。

骨干网络	语音编码器	语音解码器	词错误率（%）在语音响应
Qwen3-VL-4B	2	4	17.40
	2	6	16.40
	4	4	14.15
	6	4	12.00
	8	4	12.52
Qwen3-VL-8B	6	4	14.86
	8	4	12.49
	10	4	17.33

[7]https://huggingface.co/Qwen/Qwen2.5-3B