面向视觉语音识别的音频原生自动语音识别迁移的 CIF-查询桥接

Chang Sun[1,2], *Dongliang Xie*[1],*Wanpeng Xie*[1],*Bo Qin*[2],*Hong Yang*[2]

[1]网络与交换技术国家重点实验室，北京邮电大学[2]中华人民共和国公安部第一研究所，北京 100048，中国

arXiv:2611.32697v1 [cs.SD] 24 Nov 2026

摘要

音频原生自动语音识别（ASR）模型具备强大的语音-语言先验，但其紧凑的声学接口与较长的帧级视觉特征并不匹配。我们提出了 *CIF-Query Bridge*，这是一种从 Paraformer 视觉语音识别（VSR）编码器连接至 Qwen3-ASR 的轻量级桥接器。现有的视觉连续积分发放（CIF）预测器会生成数量取决于输入的令牌速率查询；一个两层交叉注意力桥接器利用完整帧序列对这些查询进行细化，随后由线性适配器将其输入预训练 ASR 投影器和解码器。该设计在避免对数百个视频帧进行直接自回归处理的同时，保留了对未压缩视觉上下文的访问能力。在 CMLR 上，完整的纯视觉迁移系统达到了 7.97% 的字符错误率；在相同测试集划分和评分协议下，原生 Paraformer 的字符错误率为 11.78%，前者绝对降低 3.81 个百分点，相对降低 32.3%。受控消融实验分析了桥接器深度、投影器初始化、适配器容量和 CTC 监督的影响。结果表明，基于 CIF 的自适应查询为将音频原生自动语音识别能力迁移至视觉语音识别提供了紧凑的接口。

索引词— 视觉语音识别、唇读、连续积分发放、音频原生自动语音识别、跨模态迁移

1. 引言

视觉语音识别（VSR），也称为唇读，仅凭可见的嘴部运动转写语音。当音频不可用或受到干扰，以及进行无声或注重隐私的交互时，它具有重要价值。然而，VSR 本质上存在歧义：不同音素可能具有相似的视素，而协同发音会使某个视觉单元依赖周围语境。LipNet 等早期端到端系统展示了视觉特征提取与句子预测的联合建模 [1]。普通话带来了额外挑战，因为视觉上相似的发音可能对应不同的汉字和含义。CMLR 的提出旨在研究这种大词汇量、句子级的场景 [2]。

近期的视觉语音识别进展结合了更强的序列建模、多模态预训练和数据规模扩展。Conformer 建立了强大的监督式基线 [3]；AV-HuBERT 学习掩码音频视觉单元 [4]；VATLM 统一了视觉、音频和文本预训练 [5]；RAVEn 迁移了原始音频视频监督 [6]；Auto-AVSR 则利用自动转写扩大训练规模 [7]。这些方法学习到了强大的视觉编码器，但其任务特定解码器在解决视觉歧义方面的语言建模能力较弱。

大语言模型具备更强的语言先验。VSP-LLM 将压缩的视觉语音单元映射到大语言模型 [8]；VALLR 根据预测的音素重构文本 [9], , 而 VALLR-Pin 则增加了拼音引导的普通话细化 [10]。Llama-AVSR 使用模态投影器和 LoRA，将音视频编码器连接到通用大语言模型 [11, 12], , 而 MMS-LLaMA 则使用速率感知的 Q-Former 查询 [13]。因此，动态查询分配并非我们方法所独有；在本文中，现有的视觉 CIF 预测器同时提供查询内容和长度。与通用文本大语言模型不同，音频原生 ASR 模型需要紧凑的声学表示序列。帧级视觉特征在模态、维度和长度上均存在差异，因此，如果没有专门的连接机制，就无法直接利用 ASR 解码器的语言能力。

我们研究了如何保留音频原生自动语音识别的输入接口，并将其复用于仅视觉识别。我们的*CIF-Query Bridge* 将预训练 Paraformer VSR 编码器连接至 Qwen3-ASR [14]。我们不重新设计 VSR 模型的连续积分与触发预测器，而是复用它 [15, 16]来生成数量取决于输入的查询令牌。交叉注意力利用完整的帧级视觉记忆对这些查询进行细化，然后将其传递给一个小型适配器和原始自动语音识别投影器。这种构造保留了语音接口，并避免将数百个视频帧直接呈现给自回归解码器。

我们的贡献主要有三点。第一，我们提出了一种紧凑的 CIF-Query Bridge，用于连接预训练的 VSR 前端与 Qwen3-ASR 投影器和解码器。第二，我们将现有的 CIF 输出作为依赖输入的查询，使其能够关注完整的视觉序列，从而结合令牌速率压缩与帧级记忆。第三，我们报告了在 CMLR 上取得的 7.97% 字符错误率，并研究了桥接器深度、投影器初始化、适配器容量以及 CTC 监督在

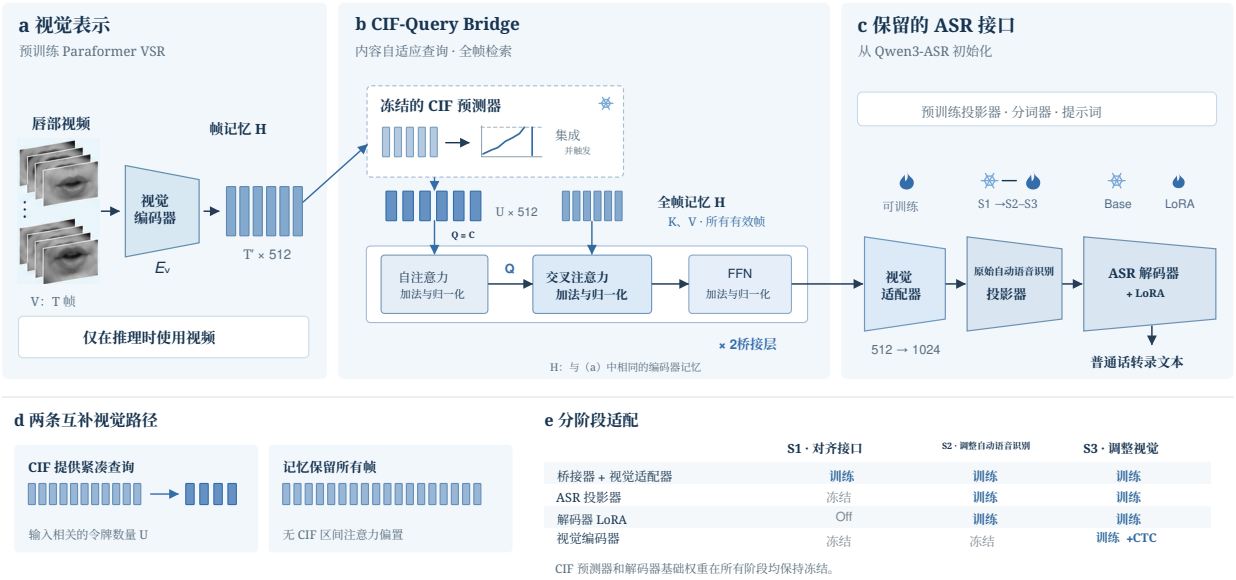

图 1。CIF-Query Bridge 概览。CIF 生成令牌速率查询 $\mathbf{C}$，而完整的视觉序列 $\mathbf{H}$ 仍作为交叉注意力记忆。主系统采用线性适配器；残差变体则通过消融实验进行评估。模块标签显示了分阶段的调度方案。

受控调度下的表现。

2. 提出的方法

2.1. 架构概览

如图 1 所示，给定一个包含 T 帧的唇部视频$\mathbf{V}$，预训练视觉编码器 E_{v} 会生成

$$\mathbf{H} = E_{\mathrm{v}}(\mathbf{V}) \in \mathbb{R}^{T' \times d_{\mathrm{v}}}, \qquad d_{\mathrm{v}} = 512, \tag{1}$$

其中通过二值掩码排除填充位置。系统将$\mathbf{H}$ 保留为帧级记忆。冻结的 CIF 预测器生成较短的序列$\mathbf{C}$；两层桥接器使用$\mathbf{C}$ 作为查询，并使用$\mathbf{H}$ 作为键和值。视觉适配器将桥接器输出映射到预期的声学特征维度。生成的特征经过原始 Qwen3-ASR 投影器，并替换其提示词中的原生语音占位符。因此，分词器、提示模板、投影器和自回归解码器仍与其原始 ASR模型配套使用。推理时不需要音频或任何源自音频的特征。

2.2. CIF 查询交叉注意力桥接器

我们采用预训练 Paraformer VSR模型中的 CIF 预测器，不改变其架构。它在每个编码器步长处估计一个权重 α_t，并对连续的加权特征进行积分，直到累积质量达到 1。这将生成

$$\mathbf{C} = [\mathbf{c}_u]_{u=1}^{U} \in \mathbb{R}^{U \times d_{\mathrm{v}}}, \qquad U = \left\lfloor \sum_{t=1}^{T'} \alpha_t + \tau \right\rfloor. \tag{2}$$

其中，$\tau = 0.45$ 是继承的尾部质量，并附加了一个零特征；实现中的保护措施会将数量限制在可用输出长度范围内，且至少保留一个位置。适配和推理都使用预测长度，不进行参考长度缩放。CIF 提供查询内容和数量，而非 Qwen 文本的词元边界：原始预测器使用字符加 EOS 长度目标进行训练。其输出作为查询，可访问未压缩视觉记忆。

令$\mathbf{Z}^0 = \mathbf{C}$。桥接层 l 依次应用词元自注意力、与视觉记忆的交叉注意力以及前馈网络：

$$\widetilde{\mathbf{Z}}^l = \mathrm{LN}\left(\mathbf{Z}^{l-1} + \mathrm{MHSA}(\mathbf{Z}^{l-1})\right), \tag{3}$$

$$\widehat{\mathbf{Z}}^l = \mathrm{LN}\left(\widetilde{\mathbf{Z}}^l + \mathrm{MHA}(\widetilde{\mathbf{Z}}^l, \mathbf{H}, \mathbf{H})\right), \tag{4}$$

$$\mathbf{Z}^l = \mathrm{LN}\left(\widehat{\mathbf{Z}}^l + \mathrm{FFN}(\widehat{\mathbf{Z}}^l)\right). \tag{5}$$

我们使用两层、八个注意力头、GELU 激活函数和四倍前馈网络扩展，并设置丢弃率为 0.1（公式中省略）。我们不使用 CIF 区间距离偏置。预测器参数保持冻结，但视觉编码器更新时，其预测的积分权重和触发边界可能会发生变化。

2.3. 视觉到自动语音识别接口投影

桥接器生成 512 维视觉标记，而 Qwen3-ASR 投影器需要 1024 维声学特征。我们通过视觉适配器 G_{v} 解决这一维度不匹配问题。其最简形式是线性映射，

$$\mathbf{a} = W_{\mathrm{v}}\mathbf{Z}^2 + \mathbf{b}_{\mathrm{v}}, \tag{6}$$

随后连接保留的两层 ASR 投影器（下文省略偏置项）

$$\mathbf{A} = W_2\,\mathrm{GELU}(W_1\mathbf{a}). \tag{7}$$

主系统采用这一线性适配器。消融实验还考察了门控残差多层感知器（$1024 \to 2048 \to 1024$，丢弃率 0.1），其 Sigmoid 门控使用 $g = -2$ 进行初始化。

2.4. 目标函数与分阶段适配

自回归解码器通过教师强制交叉熵 $\mathcal{L}_{\mathrm{AR}}$进行优化；提示标记和插入的视觉位置均不参与监督。在最终的视觉编码器适配期间，其输出还会接受 CTC 监督：

$$\mathcal{L} = \mathcal{L}_{\mathrm{AR}} + \lambda_{\mathrm{CTC}}\mathcal{L}_{\mathrm{CTC}}, \qquad \lambda_{\mathrm{CTC}} = 0.1. \tag{8}$$

不使用 CIF 数量损失。冻结预测器参数并不能保证视觉编码器适配期间输出长度保持不变。

优化分三个阶段进行。阶段1仅训练桥接器和视觉适配器，同时冻结两个预训练骨干网络以及 ASR 投影器。阶段2进一步启用原始投影器，以及语言解码器中的秩为16的 LoRA 模块。阶段3以小得多的学习率对所有视觉编码器模块进行短暂适配，并启用 CTC。这一顺序首先建立跨模态接口，然后再修改任一预训练骨干网络。

3. 实验

3.1. 数据集与实现

我们使用官方 CMLR 划分（61小时、11名说话人；训练集/验证集/测试集分别包含71,448/10,206/20,418条语句） [2],灰度 88×88 嘴部裁剪图像，以及 CER。我们报告的所有输入均为纯视觉输入。语料库 CER 是字符编辑距离总和除以参考文本长度总和；验证集和测试集分别包含145,832和292,180个参考字符。遵循 M2MeT，参考文本会针对空格、标注、标点符号和拉丁字符进行规范化，而假设文本仅移除特殊标记以及 ASR/思考前缀。

50 层 Paraformer 编码器/CIF 预测器和 Qwen3-ASR-1.7B 初始化视觉和语言组件。训练使用 AdamW，权重衰减为零，在四块 GPU 上使用 bfloat16、两步梯度累积、裁剪阈值为 1.0、余弦衰减至峰值学习率的 0.05，并将每块 GPU 的批次预算设为 500 帧。阶段 1–3 均运行 $8 + 4 + 1$ 个训练周期。阶段 1 使用桥接器 / 适配器学习率 2×10^{-4}，并进行 1,000 次预热更新。阶段 2 为桥接器、适配器、投影器和 LoRA 使用学习率 2×10^{-5}，并进行 400 次预热更新。阶段 3 分别为编码器、CTC 头、LoRA 以及其余可训练模块使用学习率 10^{-7}、 10^{-6}、10^{-5}和 5×10^{-6}，并进行 400 次预热更新；该阶段执行八周期余弦计划中的一个周期。秩为 16 的 LoRA 使用缩放系数 32、丢弃率 0.05，以及 q/k/v/o 加上 gate/up/down 投影。亮度、对比度、伽马和分区均值时间掩码仅应用于阶段 3。Beam-8 解码最多生成 64 个词元，不使用采样或 CTC 重评分。

受控变体使用随机种子 42、固定调度方案和阶段结束检查点。验证集选择双层线性模型，并从 $\{10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$ 中选择随机投影器的第一阶段学习率；其后续阶段与预训练投影器运行保持一致。无 CTC 运行从残差模型的第二阶段检查点分支而来。

3.2. 与先前工作的比较

表 1 比较了先前工作与我们的原生基线。“外部”表示超出 61 小时 CMLR 训练集的知识。我们的视觉编码器和 CIF 预测器仅使用 CMLR。唯一的外部来源是 Qwen3-ASR，其音频编码器未被使用。仅使用 CMLR 的已列出最佳结果为 9.10% 字符错误率；经过 1,459 小时的音视频预训练后，结果达到 8.00%。我们的 7.97% 并非在数据匹配条件下声称优于其他方法。

原生 Paraformer 基线使用相同的视觉初始化检查点和其原始解码器，并采用相同的视频预处理和评分方案在同一测试集上进行评估。其波束大小为 1 时，字符错误率为 11.78%；相比之下，我们的迁移系统在波束大小为 8 时，字符错误率为 7.97%。绝对降幅为 3.81 个百分点（相对降幅约为 32.3%）。该比较衡量的是完整迁移系统的表现，包括解码器、适配和搜索方面的变更；它并未单独评估 CIF-Query Bridge。

Qwen3-ASR 报告称，AuT 使用了 4000 万小时的数据，Omni 预训练使用了 3 万亿个词元 [14]。这两个阶段彼此独立，其语料库清单以及与 CMLR 的重叠情况均未知；因此，我们的结果衡量的是基础模型迁移，而非数据匹配的比较。

3.3. 消融研究

表2报告了固定调度下的验证集和测试集字符错误率。为公平地对照残差模型，随机投影器从第一阶段更新开始训练；验证集从 2×10^{-4} 中选择 $\{10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$，第二至第三阶段与预训练运行保持一致。预训练将测试集字符错误率从8.2196%降至7.9923%（配对 95% 置信区间 $[-0.3270, -0.1282]$）。

表1。官方 CMLR测试集字符错误率（%）；所有系统均使用仅视觉推理。外部：CMLR 之外的预训练；音频视觉：音频视觉。

方法	外部	CMLRh 字符错误率 ↓	
CSSMCM [2]	–	61	32.48
LIBS [17]	自动语音识别教师	61	31.27
CTCH [18]	–	61	22.01
JEP-KD [19]	自动语音识别教师	61	14.26
Ma等人 [20]	–	61	9.10
Ma等人 [20]	818-h 音频视觉	61	8.10
Ma等人 [20]	1459 小时音频视觉	61	8.00
原生 Paraformer（我们的模型）	–	61	11.78
我们的模型（线性适配器）	Qwen3-ASR	61	**7.97**

表 2。受控 CMLR 消融实验，CER（%）。Proj.：投影器初始化；Res.：门控残差适配器。所有行均使用一个训练随机种子。

投影器	层	适配器	CTC	Val	Test
随机	2	Res.	Yes	8.1464	8.2196
预训练	0	线性	Yes	8.1210	8.0820
预训练	1	线性	Yes	8.0140	7.9971
预训练	2	线性	Yes	**7.8858**	**7.9721**
预训练	2	Res.	Yes	8.0702	7.9923
预训练	2	Res.	No	8.1018	8.0776

采用线性适配器时，深度从0增至2，验证集/测试集字符错误率单调下降0.2352/0.1099个百分点；测试集置信区间 $[-0.223, +0.005]$ 包含零。线性模型与残差模型相比，测试集表现没有明显优势（-0.0202；$[-0.106, +0.069]$）；而CTC使残差模型的测试集字符错误率降低了0.0852个百分点（$[-0.154, -0.014]$）。图2报告了固定模型的配对话语级自助法置信区间；这些区间未考虑随机种子变异，且验证集结果仍用于模型选择，而非作为独立证据。

4. 讨论

完整系统获得的 3.81 个百分点提升，主要可能反映了 Qwen3-ASR 解码器及其语言先验的作用。但这并不意味着连接本身处于次要地位：自动语音识别后端需要的是一段由 1024 维声学表示构成的短序列，而视觉前端输出的则是更长的 512 维特征序列。如果没有自适应速率压缩、维度对齐以及对帧级记忆的访问，解码器的预训练能力就无法应用于视觉输入。我们的贡献在于这一连接，而不是解码器本身。此外，原生对比同时改变了解码器、适配方式和波束搜索，因此它衡量的是完整迁移系统，无法单独分离桥接器的因果贡献。

With在受控运行中，两个数据划分上的单调深度 t rend

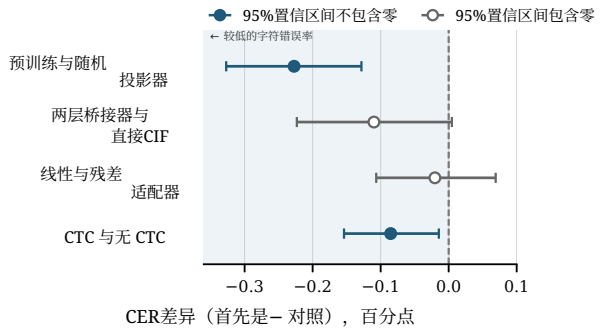

图 2。配对测试集字符错误率差异及 95% 的话语级自助法置信区间（10,000 个样本）；负值表示前一种配置更优。

与全记忆检索恢复了令牌速率 CIF集成可能丢弃的上下文这一点是一致的。不过，测试集区间仍包含零，而且深度也会改变容量和优化过程。要确定这一趋势的来源，还需要使用等长池化查询或学习型查询，以及容量匹配的无记忆对照。尽管 CIF 参数保持冻结，阶段3仍会改变其编码器输入，并可能导致触发权重、边界和预测长度发生偏移；测量这种漂移及其与字符错误率的关系，是另一项有价值的测试。

目前的证据仅限于一个随机种子、CMLR 以及一组 Paraformer–Qwen3-ASR 配对。历史测试反馈削弱了独立验证的力度，而基础模型预训练与 CMLR 之间的重叠程度尚不明确。要确立稳定性和泛化能力，还需要使用多个随机种子、匹配的文本大语言模型基线以及独立数据集进行验证。

5. 结论

我们提出了 CIF-Query Bridge，它在原有音频原生自动语音识别接口之前，利用自适应 CIF 输出作为令牌速率查询，对完整的视觉记忆进行检索。将内容自适应压缩与全记忆检索相结合，可以避免为每个视频帧输入一个 LLM 令牌，同时保留细粒度的时间证据。它还保留了模型特定的分词器、提示词和投影器，而不是学习一个无关的文本 LLM 输入空间。由此得到的纯视觉 Paraformer–Qwen3-ASR 系统在 CMLR 上取得了 7.97% 的字符错误率，相比之下，在相同数据划分和评分协议下，原生 Paraformer 的字符错误率为 11.78%。受控实验结果显示，桥接器深度呈现一致的趋势，并界定了辅助组件较小的影响。这些结果证明了该连接的可行性；未来工作应加入多个随机种子、匹配查询对照实验，并在未见过的数据集和语言上进行评估。

6. 参考文献

[1]Yannis M. Assael、Brendan Shillingford、Shimon Whiteson 和 Nando de Freitas，《LipNet：端到端句子级唇读》，*arXiv* 预印本 *arXiv:1611.01599*，2016年。

[2]Ya Zhao、Rui Xu 和 Mingli Song，《用于中文普通话唇读的级联序列到序列模型》，载于 *ACM Multimedia Asia* 论文集，2019年，第 1–6 页。

[3] Pingchuan Ma、Stavros Petridis 和 Maja Pantic，《基于 Conformer 的端到端视听语音识别》，载于 *IEEE* 声学、语音与信号处理国际会议论文集，2021年，第 7613–7617 页。

[4] BowenShi、Wei-Ning Hsu、Kushal Lakhotia 和 Abdelrahman Mohamed，《通过掩码多模态聚类预测学习视听语音表示》，载于 国际学习表征会议论文集，2022年。

[5] Qiushi Zhu、Long Zhou、Ziqiang Zhang、Shujie Liu、Binxing Jiao、JieZhang、Lirong Dai、DaxinJiang、Jinyu Li 和 Furu Wei，《VATLM：采用统一掩码预测进行语音表示学习的视觉-音频-文本预训练》，*IEEE* 多媒体汇刊，第 26 卷，第 1055–1064 页，2024。

[6] Alexandros Haliassos、Pingchuan Ma、Rodrigo Mira、Stavros Petridis 和 Maja Pantic，“从原始数据中联合学习视觉和听觉语音表征”，载于*IEEE/CVF*计算机视觉国际会议论文集，2023年，第 10006–10017 页。

[7]Pingchuan Ma、Alexandros Haliassos、Adriana Fernandez-Lopez、Honglie Chen、Stavros Petridis 和 Maja Pantic，“Auto-AVSR：使用自动标签的视听语音识别”，载于*IEEE*声学、语音与信号处理国际会议论文集，2023年，第 1–5 页。

[8]Jeong Hun Yeo、Seunghee Han、Minsu Kim 和 Yong Man Ro，“当视觉语音遇上语言：用于高效且具备上下文感知能力的视觉语音处理的 VSP-LLM 框架”，载于计算语言学协会论文集：*EMNLP 2024* 的研究成果，2024年，第 11391–11406 页。

[9] Marshall Thomas、Edward Fish 和 Richard Bowden，《VALLR：用于唇读的视觉自动语音识别语言模型》，*arXiv* 预印本 *arXiv:2503.21408*，2025年。

[10] Chang Sun、Dongliang Xie、Bo Qin 和 Hong Yang，“VALLR-Pin：通过拼音引导的大语言模型优化实现普通话双解码视觉语音识别”，*arXiv* 预印本 *arXiv:2512.20032*，2025年。

[11]Umberto Cappellazzo、Minsu Kim、Honglie Chen、Pingchuan Ma、Stavros Petridis、Daniele Falavigna、Alessio Brutti 和 Maja Pantic，“大语言模型是强大的视听语音识别学习器”，载于 *IEEE*声学、语音与信号处理国际会议论文集，2025年，第1–5页。

[12] Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen，《LoRA：大语言模型的低秩适配》，发表于国际学习表征会议论文集，2022年。

[13] Jeong Hun Yeo、Hyeongseop Rha、Se Jin Park 和 Yong Man Ro，《MMS-LLaMA：利用最少的多模态语音标记实现高效的基于大语言模型的音视频语音识别》，发表于计算语言学协会论文集：*ACL 2025*，2025年，第20724–20735页。

[14] Xian Shi、Xiong Wang、Zhifang Guo、Yongqi Wang、Pei Zhang、Xinyu Zhang、Zishan Guo、Hongkun Hao、Yu Xi、Baosong Yang、Jin Xu、Jingren Zhou 和 Jun-yang Lin，《Qwen3-ASR 技术报告》，*arXiv* 预印本 *arXiv:2601.21337*，2026年。

[15] Linhao Dong 和 Bo Xu，《CIF：用于端到端语音识别的连续积分发放》，载于*IEEE*声学、语音与信号处理国际会议论文集，2020年，第6079–6083页。

[16] Zhifu Gao、Shiliang Zhang、Ian McLoughlin 和 Zhi- jie Yan，《Paraformer：用于非自回归端到端语音识别的快速精准并行变换器》，载于*Interspeech*论文集，2022年，第2063–2067页。

[17] Ya Zhao、Rui Xu、Xinchao Wang、Peng Hou、Haihong Tang 和 Mingli Song，《听见嘴唇：通过蒸馏语音识别器改进唇读》，载于*AAAI*人工智能会议论文集，2020年，第34卷，第6917–6924页。

[18]Shihui Ma、Shilin Wang 和 Xiang Lin，《用于句子级普通话唇读的基于变换器的模型》，载于*IEEE*网络空间数据科学国际会议论文集，2020年，第78–81页。

[19] Chang Sun、Hong Yang 和 Bo Qin，“JEP-KD：用于视觉语音识别的联合嵌入预测架构知识蒸馏”，*IEEE* 开放信号处理期刊，2024。

[20] Pingchuan Ma、Stavros Petridis 和 Maja Pantic，“自然环境下的多种语言视觉语音识别”，《自然·机器智能》，第 4 卷，第 11 期，第 930–939 页，2022年。