基于大语言模型的鲁棒音视频语音识别：稀疏模态对齐与视觉单元引导细化

Fei Su [1,2], *Cancan Li*[1,2], *Juan Liu* [2,1,**], *WeiJu*[4], *HongbinSuo*[4], *MingLi* [3,5]

[**1] 中国武汉大学计算机科学学院[2] 中国武汉大学人工智能学院[3] 中国香港中文大学（深圳）人工智能学院[4] 中国OPPO人工智能中心[5] 中国昆山杜克大学数字创新研究中心fei.su@whu.edu.cn, cancan.li@whu.edu.cn, liujuan@whu.edu.cn, juwei@oppo.com, suohongbin@oppo.com, ming.li.cuhksz@gmail.com

arXiv:2603.03811v1 [cs.SD] 4 Mar 2026

摘要

音频-视觉语音识别（AVSR）融合声学和视觉信息，以提升在不利声学条件下的鲁棒性。近年来，大语言模型（LLM）取得了具有竞争力的自动语音识别性能，并展现出应用于AVSR的有效性。然而，先前的方法独立映射音频和视觉特征，或采用浅层融合，限制了跨模态对齐与互补信息交换，同时增加了大语言模型的计算负担。为此，我们提出AVUR-LLM，这是一种基于大语言模型的音频-视觉语音识别方法，通过稀疏模态**对**齐和**视**觉**单**元引导**细**化实现。LRS3上的实验表明，该方法在AVSR任务中取得了最先进的结果。在0dB信噪比的加性噪声条件下，其相较于基线系统实现了37%的相对提升。

索引词：视听语音识别、视觉离散单元、大语言模型

1. 引言

音频-视觉语音识别（AVSR）融合声学信号与视觉线索，以提升不利声学条件下的鲁棒性和准确率。早期的监督式流程涵盖 DNN/CTC [1, 2], 序列到序列 [3], 以及 RNN-T 方案 [4], , 为后续研究奠定了基础，但仍容易受到噪声影响。自监督学习通过预训练高质量语音和音视频编码器，提升了数据效率和迁移能力。wav2vec [5]和 WavLM [6] 等自监督声学模型提升了表征质量。采用 AV-HuBERT [7] 和 AV-data2vec [8] 进行音视频预训练，进一步提升了音视频语音识别的性能，而自动标签训练则带来了额外收益 [9]。近年来，大语言模型（LLM）已被证明对语音识别有效；相关研究将其作为参数化语言模型，或与语音编码器结合用作多模态推理器，取得了具有竞争力的识别性能 [10]。在音视频语音识别领域，近期基于大语言模型的框架将 Whisper [11] 与视觉特征 [12] 融合，并将 Whisper 与 LLaMA 的不同变体结合，以利用多模态上下文 [13, 14]。

以往关于基于大语言模型的视听语音识别的研究 [15, 16, 17] 通常使用自监督编码器作为特征提取器。声学特征和视觉特征在时间维度上进行压缩，并线性投影到大语言模型嵌入空间中，随后利用大语言模型进行预测或浅层融合。基于大语言模型的另一种方法语音识别采用重评分策略 [18, 19, 20, 21]，，将声学特征提供给大语言模型，或应用纯文本重评分来减少识别错误。然而，尽管取得了这些进展，基于大语言模型的语音识别仍面临一些局限：与声学/视觉特征的紧密耦合增加了内存消耗，并提高了对输入噪声的敏感性。此外，多模态融合缺乏细粒度控制。为降低计算成本并提升鲁棒性，近期研究 [22, 23, 24]通过向量量化将音频或视觉嵌入离散化，生成能够抑制高频变化并缩短大语言模型提示的令牌序列。在这些工作的基础上，我们提出了一种轻量且稳定的对齐机制，使跨模态交互能够受到控制，同时不干扰预训练音频通路。

受上述工作的启发，我们进一步探索受控的多模态交互与自适应融合，并利用语音信息视觉离散令牌引导基于大语言模型的重评分。本文提出 AVUR-LLM，主要贡献如下：(1) 一种面向稳定性的稀疏对齐方法，将轻量级跨模态模块插入音频编码器的高层；(2) 一种置信度感知融合方法，用于调节令牌级视觉注入，并采用基于视觉离散单元的提示进行大语言模型重评分；(3) 在 LRS3 数据集上的实验表明，我们的模型优于音视频语音识别基线系统，并在干净和噪声条件下均降低了词错误率，在 0 dB 信噪比下相对降低 37%。

2. 方法

我们提出的框架如图 1 所示。该方法包括稀疏模态对齐、自适应调制融合和视觉单元引导细化。

2.1. 稀疏模态对齐

我们提出了稀疏模态对齐（SMA），将轻量级对齐模块稀疏地插入上层音频编码器中。在每个模块中，视觉特征作为查询，关注音频上下文，从而校准视觉表征。

我们采用 Whisper 编码器 [11] 处理音频，采用 AV-HuBERT 编码器 [7] 处理视觉输入，分别提取特征 $\mathbf{X}_a \in \mathbb{R}^{T_a \times D_a}$ 和 $\mathbf{X}_v \in \mathbb{R}^{T_v \times D_v}$。随后，我们利用音频通路更高的时间保真度和强大的预训练能力，引导视觉对齐与增强。将视觉特征对齐到音频的时间分辨率，可缓解帧率不匹配。将对齐限制在音频编码器的上层

**表示通讯作者。

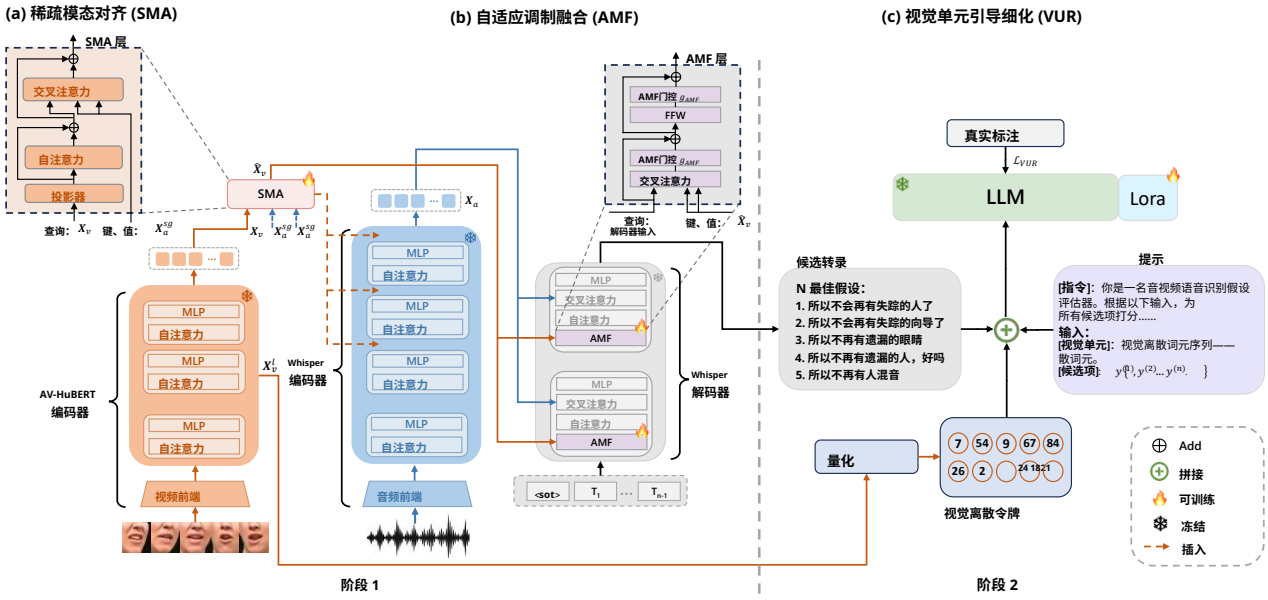


图 1： *AVUR-LLM* 示意图。*(a)* 稀疏模态对齐（*SMA*）模块：以音频为条件的交叉注意力，稀疏地插入音频编码器的高层；*(b)* 自适应调制融合（*AMF*）模块：感知置信度的门控，自适应调节解码器中的视觉注入；*(c)* 视觉单元引导细化（*VUR*）模块：将中间层视觉特征离散化为词元，供大语言模型重新评分。训练分为两个阶段：阶段 1 执行 *SMA+AMF*，并输出 N 个最佳假设；阶段 2 提取并压缩 X_v^{ℓ} 为视觉离散令牌，并使用经过 *LoRA* 适配的大语言模型对假设重新评分。这里，$\mathbf{X}_a$ 和 $\mathbf{X}_v$ 表示音频和视觉编码器特征。$\hat{}\mathbf{X}_v$ 表示经过 *SMA* 细化的视觉特征，$\mathbf{X}_v^{\ell}$ 表示从编码器第 ℓ 层提取的视觉特征。$\mathbf{X}_a^{\mathrm{sg}}$表示作为键/值且应用停止梯度的音频特征。 g_{AMF} 表示 *AMF* 门控系数。

并对音频键/值应用停止梯度，可保留声学表征、稳定优化，并且只带来适度的计算开销。

具体而言，我们保留音频分辨率，同时通过轻量级可微重采样器对视觉特征进行上采样，并将其投影至音频特征维度，从而得到$\mathbf{X}'_v = \mathcal{R}_v(\mathbf{X}_v)\,W_v + b_v$，其中$\mathbf{X}'_v \in \mathbb{R}^{T_a \times D_a}$。此处，$\mathcal{R}_v(\cdot)$表示轻量级可微重采样，它将视觉特征对齐至时间分辨率 T_a。 $W_v \in \mathbb{R}^{D_v \times D_a}$和 $b_v \in \mathbb{R}^{D_a}$ 是线性投影器的可学习参数，用于将特征维度从 D_v 转换为 D_a。

随后，我们在音频编码器的上层插入轻量级模块，在保持音频特征不变的同时，使视觉特征与音频线索对齐。设$\mathbf{X}_a \in \mathbb{R}^{T \times D}$ 和$\mathbf{X}'_v \in \mathbb{R}^{T \times D}$，其中 $T = T_a$ 和 $D = D_a$。每个模块：

$$\begin{aligned} \mathbf{X}_v^{\mathrm{sa}} &= \mathrm{MHA}(\mathbf{X}'_v, \mathbf{X}'_v, \mathbf{X}'_v), \\ \hat{\mathbf{X}}_v &= \mathbf{X}'_v + \mathrm{MHA}(\mathbf{X}_v^{\mathrm{sa}}, \mathbf{X}_a^{\mathrm{sg}}, \mathbf{X}_a^{\mathrm{sg}}), \end{aligned} \tag{1}$$

其中，MHA表示多头注意力， [25], sa表示经过自注意力处理的视觉特征，sg表示停止梯度，确保音频特征仅作为键/值且保持不变。我们在音频编码器的上层插入三个对齐模块，使对齐聚焦于高层表示，同时保持计算效率和音频编码器稳定性。

2.2. 自适应调制融合

为了在解码过程中自适应地利用互补视觉线索，我们首先通过仅前向探测器，基于解码器输入和音频键/值估计词元级声学可靠性信号。随后，我们利用该信号控制视觉注入的幅度。受 [12], 启发，我们还采用基于双曲正切的方向门控，以确定注入的视觉特征发挥正向还是负向作用。

具体而言，我们通过在每一层插入一个仅前向声学探针来调整 Whisper 的解码器，并停止梯度通过查询和音频特征的传播。给定解码器查询$\mathbf{Q}_{\ell} \in \mathbb{R}^{T_y \times D}$ 和音频记忆 $\mathbf{X}_a \in \mathbb{R}^{T \times D}$，探针注意力为：

$$A_{\ell} = \mathrm{softmax}\left(\frac{(\mathbf{Q}_{\ell}^{\mathrm{sg}} W_Q)\,(\mathbf{X}_a^{\mathrm{sg}} W_K)^{\top}}{\sqrt{D}}\right). \tag{2}$$

词元级声学不确定性 $S_{\ell,t}$ 通过计算探针注意力在音频帧 A_{ℓ}^{t} 上的熵，并以 log T归一化为 $[0,1]$ 得到。$S_{\ell,t}$ 的值表示注意力在音频帧上的集中程度，其中较小的值对应更可靠的声学上下文，而较大的值反映更高的声学不确定性。因此，视觉特征注入的幅度定义为：

$$g_{\mathrm{amp},t} = \sigma\big(a\,S_{\ell,t} + b\big), \tag{3}$$

其中， $a, b \in \mathbb{R}$ 是用于控制斜率和偏移量的可学习标量， $\sigma(\cdot)$ 表示 Sigmoid 函数，确保 $g_{\mathrm{amp},t} \in (0,1)$。

给定解码器输入 $\mathbf{x}_t$，通过以 LN($\mathbf{x}_t$) 为查询、以 $\hat{}\mathbf{X}_v$ 为键/值的 MHA 得到视觉上下文$\mathbf{c}_t^{(v)}$ 。自适应调制融合（AMF）门控在每层使用两个标量：$g_{\mathrm{AMF,att}}$ 用于交叉注意力分支，$g_{\mathrm{AMF,ff}}$ 用于前馈分支。每个门控都结合了一个方向项（将 tanh(·) 应用于可学习标量得到）和一个逐词元的幅度项 $\sigma(\cdot)$，该幅度项由声学不确定性 $S_{\ell,t}$ 推导得出。AMF 门控分别调制交叉注意力和前馈

前向残差：

$$\begin{aligned}\mathbf{x}'_t &= \mathbf{x}_t + g_{\mathrm{AMF,att}}\,\mathbf{c}_t^{(v)},\\ \mathbf{y}_t &= \mathbf{x}'_t + g_{\mathrm{AMF,ff}}\,\mathrm{FFW}\big(\mathrm{LN}(\mathbf{x}'_t)\big),\end{aligned} \tag{4}$$

其中，LN 表示层归一化，FFW 表示多层感知机（MLP）。我们使用模型预测的转录结果与真实词元之间的令牌级交叉熵进行训练。在优化过程中，可训练组件包括稀疏模态对齐模块和自适应调制融合模块。在推理阶段，束搜索生成一个 N最佳列表 $Y=\{y^{(i)}\}_{i=1}^{N}$，其序列分数为 $s_{\mathrm{infer}}\big(y^{(i)}\big)$.

2.3. 视觉单元引导的优化

由于 AV-HuBERT 编码器的视觉表征包含与语音相关的结构信息，我们将其离散化为视觉令牌，并将所得序列作为条件输入提供给大语言模型（LLM），用于对 N最佳结果进行重评分。不同于以往基于 LLM 的方法 [15,16, 20]将连续音视频特征投影或融合表示输入 LLM，VUR 提供了源自 AV-HuBERT 的紧凑视觉令牌序列，作为明确的视觉提示。

具体而言，我们使用来自视觉编码器的视觉嵌入$\mathbf{X}_v^{\ell}\in\mathbb{R}^{T_v\times D_v}$，其中帧向量为$\mathbf{x}_{v,\,t}^{\ell}\in\mathbb{R}^{D_v}$。通过离线训练带有质心 $\{\mathbf{k}_1,\dots,\mathbf{k}_K\}$ 的 K-means 码本$\mathbf{k}_j$，并将每一帧分配给最近的质心，对其进行量化。为利用时间平稳性并缩短令牌序列，我们应用游程压缩。具有相同标签的最大连续片段会被合并为一个令牌：对每个片段中的帧取平均值，并保留共同标签 [13]。令 $R_1,..,R_M$ 表示具有相同标签的最大连续索引集合。对于每个 R_m，我们计算平均特征 $\bar{}\mathbf{x}_{v,m}^{\ell}=|R_m|^{-1}\sum_{t\in R_m}\mathbf{x}_{v,t}^{\ell}$，并分配共同标签 k_m。所得的视觉离散单元序列会针对每个话语缓存一次，并在 N-best 重评分中重复使用。

基于第一阶段的推理结果，我们使用 LoRA [27] 对大语言模型 [26] 进行适配，为每个候选项分配一个标量分数 r_i，其中 LoRA 参数可训练。令 i_{gt} 表示与真实转录文本最接近的候选项。训练目标是最小化列表式Softmax：

$$\mathcal{L}_{\mathrm{VUR}}=-\log\frac{\exp(r_{i_{\mathrm{gt}}})}{\sum_{j=1}^{N}\exp(r_j)}, \tag{5}$$

其中，$r_j\in\mathbb{R}$ 是大语言模型为第 j个候选项分配的分数，N 是 N优列表的大小。

3. 实验结果

3.1. 实验设置

所有实验均在 LRS3 数据集 [28], 上进行。该数据集是广泛使用的公开英语视听语音识别数据集，包含约 433 小时带转录文本的 TED演讲视频。除完整训练集外，我们还在 30 小时的“train-val”子集上评估性能。此外，按照 [12], 的方法，我们将 LRS3 训练集与 VoxCeleb2 [29]中 1326 小时的英语视频合并，以扩充训练数据。为评估噪声鲁棒性，我们按照 [12]的设置，在信噪比为 10、5、0、 -5 和 -10 dB 的条件下，添加来自 MUSAN数据集 [30] 的多人 babble 噪声。

我们使用 Whisper Medium [11] 作为音频编码器和解码器。输入特征是从 16kHz 音频中提取的 80 维对数梅尔滤波器组系数，采用 25ms 的分析窗口和 10ms 的帧移。对于视觉模态，

表1：*AVUR-LLM* 与先前自动语音识别和音视频语音识别方法在 *30* 小时、*433* 小时和 *LRS3+VoxCeleb2*（*1759* 小时）设置下于 *LRS3* 上的词错误率（%）。

方法	音频编码器	视觉编码器	训练数据（小时）	词错误率（%）↓
仅音频				
CM-seq2seq [31]	Conformer	–	433	2.3
Fast Conformer [32]	Conformer	–	435	1.6
AV-HuBERT [7]	Transformer	–	433	1.6
Whisper 微调版 [12]	Transformer	–	433	2.3
Auto-AVSR [9]	Conformer	–	1902/3448	1.0/1.0
Llama-AVSR [15]	Whisper	–	30/433/1759	1.5/1.1/0.79
AVUR-LLM	Whisper	–	30	**1.3**
AVUR-LLM	Whisper	–	433	**1.0**
AVUR-LLM	Whisper	–	1759	**0.70**
音频-视觉				
CM-seq2seq [31]	Conformer		433	2.3
AV-HuBERT [7]	Transformer		433	1.4
CMA [33]	Transformer		433	1.5
AV-data2vec [8]	Transformer		433	2.5
Fast Conformer [32]	Conformer		1687	0.9
Auto-AVSR [9]	Conformer		1902/3448	1.0/0.9
ViT3D-CM [34]	Conformer		90K	1.6
LP Conformer [35]	Conformer		100K	0.9
Whisper-Flamingo [12]	Whisper	AV-HuBERT	433/1759	1.1/0.86
Llama-AVSR [15]	Whisper	AV-HuBERT	433/1759	0.95/0.77
Llama-MTSK [17]	Whisper	AV-HuBERT	433	0.9
MMS-LLaMA [16]	Whisper	AV-HuBERT	433/1759	0.9/0.72
AVUR-LLM	Whisper	AV-HuBERT	433	**0.75**
AVUR-LLM	Whisper	AV-HuBERT	1759	**0.68**

帧级表示取自 AV-HuBERT Large [7]。在训练过程中，其权重保持冻结。对于大语言模型，我们采用 LLaMA2-7B [26]。在 LLaMA 的每一层中插入一个秩为 16 的 LoRA 模块进行微调。引入线性投影层，将视觉离散令牌映射到 LLaMA 的嵌入维度。

3.2. 训练与推理细节

SMA与AMF训练及N-best生成。视频数据以25fps采样并转换为灰度图。使用Dlib提取大小为 96×96 的唇部中心区域，并对齐到平均人脸模板。训练期间应用随机 88×88 裁剪，水平翻转概率为0.5，而推理时使用中心裁剪。对于视觉离散令牌，我们将第12层AV-HuBERT的帧级特征量化为2000个K-means聚类簇，随后进行压缩 [23]。训练采用AdamW优化器，配合余弦退火学习率调度器，权重衰减为0.1。学习率设为 1×10^{-4} 用于训练。

使用大语言模型进行 VUR。N-best假设与视觉离散令牌配对，构成大语言模型的输入。提示定义如下：［指令］：你是一名视听语音识别假设评估器。请根据以下输入，按照唇动合理性和语言连贯性为所有候选项评分。输入：［视觉单元］{**词元**}［候选项］{$y^{(1)}$,$y^{(2)}$，.....，$y^{(N)}$}，其中**词元**表示视觉离散词元序列，$y^{(1)}$，.....，$y^{(N)}$ 表示 N最佳列表中的 N 候选转录文本。使用 LoRA 微调 LLaMA2-7B，秩设为 16，dropout 率设为 0.05，并对查询、键、值和输出投影层进行修改。训练时采用的学习率为 5×10^{-4}。使用词错误率（WER）评估模型在自动语音识别和音视频语音识别任务中的表现。

表2： *LRS3*（*433*小时设置）在干净和噪声条件下的词错误率（%）。我们报告了信噪比为 $\{10, 5, 0, -5, -10\}$ *dB* 时的音视频语音识别结果。“–”表示未报告。

方法	模态	干净	信噪比等级（dB），词错误率（%）↓				
			10	5	0	-5	-10
AV-HuBERT [7]	AV	1.4	2.0	2.6	5.8	16.6	34.9
CMA [33]		1.5	1.8	2.4	4.4	11.9	25.8
Whisper-Flamingo [12]		1.1	1.4	2.0	6.3	28.0	42.6
AVGER [20]		1.1	–	–	1.9	3.7	12.3	29.1
Llama-AVSR [15]		0.95	–	–	2.2	4.2	16.9	–
MMS-LLaMA [16]		0.9	–	–	**1.3**	2.7	7.4	–
AVUR-LLM	AV	**0.75**	**1.0**	1.4	**1.7**	**6.3**	**12.9**

3.3. 实验结果与分析

3.3.1. 主要结果

表1展示了所提出方法与先前自动语音识别和音视频语音识别方法的性能对比。在纯音频设置下，AVUR-LLM 的词错误率分别为 1.3%（30 小时设置）和 1.0%（433 小时设置），优于 Llama-AVSR [15]（1.5%/1.1%）和 Whisper-Flamingo [12]（433 小时设置下为 2.3%）。在扩展监督设置下（LRS3+VoxCeleb2，1759 小时），AVUR-LLM 进一步将纯音频词错误率降至 0.70%，在相同规模下优于 Llama-AVSR（0.79%）。在音视频设置下，AVUR-LLM 在 LRS3（433 小时设置）下的词错误率为 0.75%，优于近期基于大语言模型的系统 [15, 16, 17]，并在基于大语言模型的音视频语音识别模型中创下新的最先进水平；与 Whisper-Flamingo（1.1%）和此前最佳的基于大语言模型的结果（0.9%）相比，相对词错误率分别降低了 31.8% 和 16.7%。在 1759 小时设置下，AVUR-LLM 的词错误率达到 0.68%，持续优于此前基于大语言模型的系统。值得注意的是，在 433 小时设置下，我们的结果仍能与使用更大规模监督语料库训练的 Conformer/Transformer 系统 [32] 相媲美。

3.3.2. 噪声鲁棒性评估

在表2中，我们评估了 AVUR-LLM 在不同声学噪声水平下的鲁棒性。AVUR-LLM 在大多数信噪比条件下都取得了最低的词错误率。具体而言，在 0 dB 信噪比下，AVUR-LLM 优于 MMS-LLaMA（1.7% 对比 2.7%）。在噪声条件（5/0/−5dB）下，AVUR-LLM 的总体词错误率相对降低了 14.7%。在相对干净的条件下，解码器声学注意力仍高度集中，产生较低的不确定性，并限制视觉注入。然而，随着信噪比降低，声学注意力会变得分散。AMF 会提高视觉贡献，从而显著降低错误率。在 -5 和 -10 dB 下，AVUR-LLM 将词错误率分别降至 6.3% 和 12.9%，优于此前的最佳结果 7.4% 和 25.8%，表明所提出的方法有效

在严重噪声环境下。

3.3.3. 消融结果

表3展示了消融研究的结果。启用所有组件后，词错误率从1.30%降至0.75%（干净条件下），并从0 dB 信噪比下的6.70%降至1.70%，相对降幅分别为42.3%和74.6%。移除VUR会显著降低鲁棒性（从1.70%升至5.50%），证实了细化模块与置信度感知融合结合时的有效性。在不使用SMA的情况下采用AMF+VUR，已经带来了显著提升（2.10%）；而仅使用SMA的设置（不含AMF和VUR）在噪声条件下提升甚微（6.70%），这表明置信度感知融合和细化模块

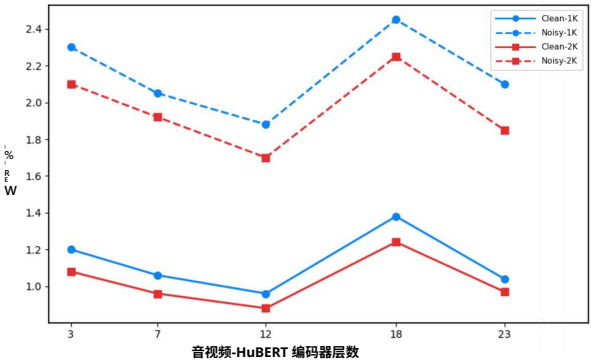


图2： 视觉离散令牌提取深度和码本大小对词错误率（%）的影响。实线表示干净条件，虚线表示 *0 dB* 信噪比。蓝色曲线表示 K均值码本大小为 $K=1000$，红色曲线表示 $K=2000$。

表3： 在干净和噪声（*0 dB* 信噪比）条件下，*AVUR-LLM* 在 *LRS3*（*433* 小时）上的消融实验。

方法	SMA	AMF	VUR	干净	含噪
Whisper-Flamingo [12]	–	–	–	1.10	6.30
Llama-AVSR [15]	–	–	–	0.95	4.20
分阶段消融					
AVUR-LLM（完整）	✓	✓	✓	**0.75**	**1.70**
– 不使用 VUR	✓	✓	✗	0.97	5.50
– 不使用 AMF（仅使用 SMA）	✓	✗	✗	1.30	6.70
– 不使用 SMA（AMF+VUR）	✗	✓	✓	0.85	2.10

是提升噪声鲁棒性的主要因素。在AMF+VUR的基础上加入SMA，可进一步将0 dB下的词错误率从2.10%降至1.70%，这表明稀疏对齐通过对齐多模态特征并缓解跨模态失配，带来了互补的收益。

3.3.4. 进一步分析

图 2 展示了在干净条件和 0dB 条件下，视觉离散单元提取深度与码本大小对词错误率的影响。性能在中间层（第 12 层）达到峰值：使用 2K 个聚类时，词错误率分别达到 0.75%（干净条件）和 1.70%（0 dB），始终低于 1K 设置。浅层（第 3 层）和深层（第 18 层）特征会产生更高的错误率，这表明中间层表示在视觉语音细节与不变性之间实现了平衡。增大码本粒度可通过提供更具判别性的视觉令牌来增强 VUR 的重评分效果，并在各种噪声条件下带来稳定的性能提升。

4. 结论

本文提出了 AVUR-LLM，这是一种基于大语言模型的视听语音识别框架，将多模态融合与视觉单元引导的细化相结合。该方法稳定了音频通路中的跨模态交互，在解码过程中自适应调节视觉贡献，并将视觉特征压缩为离散词元，以便使用大语言模型进行重评分。两阶段训练方案在保留预训练音频和视觉编码器优势的同时，减轻了大语言模型的负担。实验表明，该方法在干净和噪声条件下均有所提升，并通过对不同信噪比和离散化设置的分析，展现出更强的鲁棒性和数据效率。

5. 生成式 AI 使用披露

大语言模型（LLM）仅用于论文润色（例如改写和语法检查），以提升清晰度和可读性。未将大语言模型用于构思、方法、实验设计、数据分析或结果解读。所有科学内容均由作者撰写并核实。

6. 参考文献

[1] K. Noda、Y. Yamaguchi、K. Nakadai、H. G. Okuno 和 T. Ogata，“ 使用深度学习进行视听语音识别 ”，Applied intelligence，第 42 卷，第 722–737 页，2015。[2] A. Graves、S. Fern´andez 等，“连接时序分类：使用循环神经网络为未分段序列数据标注”，载于《国际机器学习大会论文集》，2006 年，第 369–376 页。[3] T. Afouras、J. S. Chung、A. Senior、O. Vinyals 和 A. Zisser[4] man，“深度视听语音识别”，IEEE模式分析与机器智能汇刊，第 44 卷，第 12 期，第 8717–8727 页，2018。[5] T. Makino、H. Liao 等，“用于视听语音识别的循环神经网络转导器”，载于《自动语音识别与理解研讨会论文集》，2019 年，第 905–912 页。[6] A. Baevski、Y. Zhou 等，“ wav2vec 2.0：语音表征自监督学习框架”，《神经信息处理系统进展》，第 33 卷，第 12 449–12 460 页，2020年。[7] S. Chen、C. Wang 等，“稳健的自监督音视频语音识别”，载于 Interspeech 论文集，2022 年，第 2118–2122 页。[8] J. Lian、A. Baevski 等，“AV-data2vec：利用上下文化目标表征进行音视频语音表征的自监督学习”，载于《自动语音识别与理解研讨会论文集》，2023 年，第 1–8 页。[9] P. Ma、A. Haliassos 等，“ Auto-AVSR：使用自动标签进行音视频语音识别”，载于 ICASSP 论文集，2023 年，第 1–5 页。[10] C. Chen、Y. Hu 等，“Hyporadise：基于大语言模型的生成式语音识别开放基线”，《神经信息处理系统进展》，第 36 卷，第 31 665–31 688 页，2023 年。[11] A. Radford、J. W. Kim 等，“通过大规模弱监督实现稳健语音识别”，载于《国际机器学习大会论文集》，2023 年，第 28 492–28 518 页。[12] A. Rouditchenko、Y. Gong 等，“ Whisper-Flamingo：将视觉特征融入 Whisper，以实现音视频语音识别与翻译”，载于 Interspeech 论文集，2024 年，第 2420–2424 页。[13] J. Yeo、S. Han 等，“视觉语音与语言相遇之处：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 EMNLP 论文集，2024 年，第 11 391–11 406 页。[14] U. Cappellazzo、M. Kim 等，“扩展并增强基于大语言模型的视听语音识别：一种稀疏投影器混合方法”，载于 Interspeech 论文集，2025 年，第 1823–1827 页。[15] U. Cappellazzo、M. Kim、H. Chen、P. Ma 等，“大语言模型是强大的音视频语音识别学习器”，载于 ICASSP 论文集，2025 年，第 1–5 页。[16] J. H. Yeo、H. Rha 等，“MMS-LLaMA：使用最少多模态语音词元的高效基于大语言模型的音视频语音识别”，载于《计算语言学协会年会论文集》，2025 年，第 20 724–20 735 页。[17] U. Cappellazzo、M. Kim 等，“通过基于套娃结构的多模态大语言模型实现自适应视听语音识别”，ArXiv，第 abs/2503.06362 卷，2025 年。[18] S. Radhakrishnan、C.-H. H. Yang 等，“Whispering LLaMA：一种用于语音识别的跨模态生成式纠错框架”，载于《计算语言学协会年会论文集》，2023 年，第 10 007–10 016 页。

[19] C. Chen、R. Li 等，“永远不嫌晚：将声学信息融合到大语言模型中以实现自动语音识别”，载于 ICLR 论文集，2024 年，第 437–453 页。[20] R. Liu、H. Yuan、G. Gao 和 H. Li，“再听再看：用于视听语音识别的生成式纠错”，《信息融合》，第 120 卷，第 103077 页，2025 年。[21] S. Ghosh、S. Kumar 等，“Lipger：视觉条件生成式纠错，实现稳健的自动语音识别”，载于 Interspeech 论文集，2024 年，第 1920–1924 页。[22] B. Shi、W.-N. Hsu、K. Lakhotia 和 A. Mohamed，“通过掩码多模态聚类预测学习音视频语音表征”，载于 ICLR 论文集，2022 年。[23] M. Kim、J. Yeo 等，“多语言视觉语音识别的高效训练：使用离散化视觉语音表征进行预训练”，载于 ACM MM 论文集，2024 年，第 1311–1320 页。[24] P. Mousavi、J. Duret 等，“我们应如何从自监督模型中提取离散音频令牌？ ”，载于 Interspeech 论文集，2024 年，第 2554–2558 页。[25] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. Kaiser 和 I. Polosukhin，“Attention is all you need”，《神经信息处理系统进展》，第 30 卷，2017 年。[26] H. Touvron、L. Martin 等，“Llama 2：开放基础模型与微调聊天模型”，arXiv 预印本 arXiv:2307.09288，2023 年。[27] E. J. Hu、Y. Shen 等，“LoRA：大语言模型的低秩适配”，载于 ICLR 论文集，2022 年。[28] T. Afouras、J. S. Chung 和 A. Zisserman，“LRS3-TED：用于视觉语音识别的大规模数据集”，arXiv 预印本 arXiv:1809.00496，2018 年。[29] J. S. Chung、A. Nagrani 和 A. Zisserman，“VoxCeleb2：深度说话人识别”，载于 INTERSPEECH 论文集，2018 年，第 1086–1090 页。[30] D. Snyder、G. Chen 和 D. Povey，“Musan：音乐、语音和噪声语料库”，arXiv 预印本 arXiv:1510.08484，2015。[31] P. Ma、S. Petridis 和 M. Pantic，“使用 Conformer 实现端到端视听语音识别”，载于 ICASSP 论文集，2021 年，第 7613–7617 页。[32] M. Burchi、K. C. Puvvada 等，“使用混合 CTC/RNN-T 快速 Conformer 进行多语言视听语音识别”，载于 ICASSP 论文集，2024 年，第 10 211–10 215 页。[33] S. Kim、K. Jang 等，“利用跨模态注意力学习视频时序动态，实现稳健的视听语音识别”，载于 SLT 论文集，2024 年，第 447–454 页。[34] D. Serdyuk、O. Braga 和 O. Siohan，“用于单人及多人视频视听语音识别的基于 Transformer 的视频前端”，载于 Interspeech 论文集，2022 年，第 2833–2837 页。[35] O. Chang、H. Liao 等，“视觉语音识别只需 Conformer”，载于 ICASSP 论文集，2024 年，第 10 136–10 140 页。