用于语音情感识别的大型音频语言模型知识蒸馏的注意力加权中心核对齐

Qingran Yang[1, 2†]、Botao Zhao[1†]、Zuheng Kang[1], Xue Li[2], Yayun He[1], Chuhang Liu[1], Xulong Zhang[1], Junqing Peng[1], Jianzong Wang[1*] 平安科技（深圳）有限公司，深圳，中国，[2]哈尔滨工业大学，哈尔滨，中国。

arXiv:2602.01547v1 [cs.SD] 2 Feb 2026

摘要

大型音频语言模型（LALMs）的出现推动了语音情感识别（SER）的发展，但其规模限制了在资源受限环境中的部署。尽管知识蒸馏能够有效压缩大音频语言模型，但现有方法对跨模态投影模块（投影器）的蒸馏研究仍不充分，并且由于特征维度差异，通常难以实现对齐。我们提出 PL-Distill，这是一种知识蒸馏框架，将投影器级蒸馏（PDist）用于对齐音频嵌入，将 Logits 级蒸馏（LDist）用于对齐输出 logits。PDist 引入了我们提出的新方法——注意力加权中心核对齐，通过突出重要时间步并解决维度不匹配问题来实现有效对齐。同时，LDist 最小化教师和学生在音频与文本模态上的输出 logits 之间的 KL散度。在 IEMOCAP、RAVDESS 和 SAVEE 数据集上，PL-Distill 将一个拥有 8.4B 参数的教师模型压缩为紧凑的 1.1B 参数学生模型，并在所有指标上始终优于教师模型、最先进的预训练模型以及其他知识蒸馏基线方法。

索引词——大型音频语言模型、语音情感识别、知识蒸馏、注意力加权中心核对齐

1. 引言

语音情感识别（SER）技术旨在检测语音中的情绪，在心理健康、客户服务和医疗领域展现出良好的应用前景 [1, 2, 3], , 因此吸引了广泛的研究和关注。传统的语音情感识别模型通常采用经典的深度学习架构，如卷积神经网络、循环神经网络和长短期记忆网络。随着自监督学习的兴起，WavLM [4] 和 Whisper [5] 等模型通过学习具有泛化能力的音频表征来提升语音情感识别效果。这些方法在学习声学特征方面虽有效，但捕捉语音中蕴含的文本语义的能力仍然有限。这一差距限制了它们理解复杂情绪的能力，而复杂情绪往往很大程度上取决于语言内容，而不仅仅是声学线索。Qwen2-Audio [6, 7], 等大音频语言模型（LALMs）的发展，为解决语音情感识别中的这一局限开辟了新途径。通过整合大语言模型（LLMs），将声学特征与语义信息融合，这些模型能够更深入地理解复杂情绪 [8]。尽管取得了这些进展，Qwen2-Audio 等大音频语言模型的参数规模仍然庞大，例如其拥有 84 亿个参数[6], , 导致推理成本高昂，限制了它们在资源受限场景中的应用。

知识蒸馏（知识蒸馏） [9, 10] 是一种模型压缩技术，可通过将强大的教师模型中的知识迁移至轻量级学生模型，减少大音频语言模型的参数规模，使学生模型以更少的参数和更低的计算开销达到教师模型的性能。现有的大多数知识蒸馏方法仅关注文本模态，采用前向库尔贝克-莱布勒散度（前向 KL），通过最小化 KL散度来对齐教师与学生的输出 logits。反向KL散度 [11] 和詹森-香农散度 [12] 通过改进分布度量进一步优化这种对齐。近期进展已将知识蒸馏拓展至多模态大语言模型，例如 LLaVA-MoD [13] 和 LLAVA-KD [14]。LLaVA-MoD 通过混合专家 [15] 架构优化蒸馏，但增加了结构复杂度和资源消耗。相比之下，LLAVA-KD 在不改变模型结构的情况下调整蒸馏策略，有效利用多模态信息。

然而，尽管大音频语言模型是多模态大语言模型中的重要类型，其蒸馏研究却一直未受重视。此外，将针对多模态大语言模型的蒸馏方法直接应用于大音频语言模型，会面临以下几项挑战：(1) 大多数方法是为视觉任务设计的，侧重于空间依赖关系（例如，物体间的交互）[14], , 因而无法充分捕捉音频中的时间重要性。(2) 投影器通过将非文本特征映射至语言空间，对跨模态融合至关重要，但通常未纳入蒸馏范围。即使对投影器特征进行蒸馏，教师与学生输出之间的维度不匹配仍是一项挑战 [16]。

为解决这些问题，我们提出了 PL-Distill，这是一种面向语音情感识别中大音频语言模型的轻量级知识蒸馏框架。PL-Distill 由两个主要部分组成：(1) 投影器级蒸馏（PDist）：我们提出注意力加权中心核对齐（AwCKA），这是 CKA [17], 的一种扩展，通过教师模型的自注意力分数引入动态权重。所提出的 AwCKA 不仅解决了特征维度不匹配的问题，还能优先关注情感显著的时间步。(2) Logits 级蒸馏（LDist）：该部分最小化教师与学生在音频和文本模态下 Logits 之间的 KL散度，确保跨模态一致性，同时促进声学知识和文本知识的迁移。我们的贡献总结如下：

- 我们提出了 PL-Distill，这是一种新颖的框架，填补了大语言模型知识蒸馏研究领域的空白，并显著降低了在资源受限环境中的部署成本。
- 我们提出 AwCKA，用于对齐教师

本研究获得深圳-香港联合资助项目（A类）资助，项目编号：SGDX20240115103359001。

† 这些作者对本研究的贡献相同。

* 通讯作者：Jianzong Wang (jzwang@188.com)。

和学生的音频嵌入，从而帮助投影器聚焦于关键音频词元，并解决特征维度不匹配问题。

• 大量实验表明，所提出的方法可将 Qwen2-Audio 教师模型（84 亿参数）压缩 87%，并在权威的语音情感识别数据集上取得优于当前最先进模型的性能。

2. 方法

如图1所示，我们提出的 PL-Distill 包含投影器级蒸馏和 Logits 级蒸馏。我们将通过学生模型架构以及这两个蒸馏组件来介绍我们的方法。

2.1. 学生模型的架构

我们使用 Qwen2-Audio（84 亿参数）作为教师模型，而学生模型采用相同的架构，但仅有 11 亿参数，约为教师模型规模的 13%。模型的具体结构如下：

• 音频编码器：两个模型均使用 Whisper large v3 [5] 模型作为音频编码器。训练期间，我们加载教师模型的预训练参数，并冻结该模块。

• 音频投影器：两个模型均采用单层多层感知机作为音频投影器，将音频编码器提取的音频特征映射到各自的大语言模型嵌入空间，从而生成音频嵌入。

• 大语言模型（LLM）：教师模型使用 Qwen2-7B，而学生模型使用较小的 Qwen2-0.5B，该模型由 Qwen2-0.5B-Instruct 初始化，以利用其语言能力。

在多模态整合中，音频投影器输出的音频嵌入与文本嵌入拼接，形成联合表示 $H = [\mathcal{H}_a, \mathcal{H}_p, \mathcal{H}_r]$，其中 $\mathcal{H}_a$、$\mathcal{H}_p$ 和 $\mathcal{H}_r$ 分别表示音频、提示和响应嵌入 [18]。大语言模型对该表示进行处理，生成输出 logits $[Z_a, Z_p, Z_r]$，分别对应音频、提示和响应。

2.2. 投影器级蒸馏

对于一个音频样本，教师模型和学生模型均使用相同的音频编码器提取音频特征。随后，这些特征分别经各自的音频投影器处理，得到教师的音频嵌入 $\mathcal{H}_a^{(T)} \in \mathbb{R}^{L_a \times E_T}$ 和学生的音频嵌入 $\mathcal{H}_a^{(S)} \in \mathbb{R}^{L_a \times E_S}$。其中，$L_a$ 表示音频词元序列的长度，E_T 和 E_S 表示音频词元的维度，分别对应教师和学生各自的大语言模型嵌入维度。

在投影器级蒸馏（PDist）中，为了蒸馏维度不同（$E_T \neq E_S$）的 $\mathcal{H}_a^{(T)}$和 $\mathcal{H}_a^{(S)}$，我们采用中心化核对齐（CKA） [17]。CKA 是一种通过协方差结构衡量表示对齐的相似性度量，因此特别适合比较维度不同的表示，因为它关注的是统计依赖关系，而非原始特征值 [19]。我们采用 CKA 的线性核实现，这是因为先前的工作 [17, 20] 未发现使用非线性核能显著提升准确率，而线性核具有更高的计算效率，这对于扩展到大规模模型尤其有价值。

标准线性 CKA 对所有音频词元一视同仁，而对于语音情感识别（SER）等任务，情感线索往往集中在音频词元序列中的特定关键时间步。为考虑这种时间稀疏性，并使蒸馏更关注显著性更高的部分，我们为每个音频词元引入重要性权重，由此提出注意力加权中心核对齐（AwCKA）。

音频词元权重记为 $\mathbf{w} \in \mathbb{R}^{L_a \times 1}$，取自教师模型的大语言模型的最后一个自注意力层。之所以选择最后一层，是因为它反映了全面且全局的语义信息，并能提供与任务相关的更整体理解 [21]。我们将响应词元（例如“Happy”）到音频词元的注意力分数提取为 $\mathbf{w}$ 的原始值。这些基于 Transformer 的位置编码和自注意力机制得到的注意力分数，能够自然地捕捉长期时间依赖。注意力分数越高，表示相应音频词元与响应的相关性越强，并且在传达情感方面发挥着更关键的作用。最后，我们对这些注意力分数进行归一化，得到音频词元权重 $\mathbf{w}$，如下所示：

$$\mathbf{w} = \frac{\mathbf{A}}{\sum_{i=1}^{L_a} A_i}, \tag{1}$$

其中，$\mathbf{A} \in \mathbb{R}^{L_a \times 1}$ 表示注意力分数，而 $\sum_{i=1}^{L_a} A_i$ 是所有注意力分数的总和。

然后，我们将音频词元权重 $\mathbf{w}$ 应用于教师和学生的音频嵌入 $\mathcal{H}_a^{(T)}$和 $\mathcal{H}_a^{(S)}$，得到两者的加权嵌入：$H_T \in \mathbb{R}^{L_a \times E_T}$ 和 $H_S \in \mathbb{R}^{L_a \times E_S}$。

$$H_T = \mathbf{w} \odot \mathcal{H}_a^{(T)}, \quad H_S = \mathbf{w} \odot \mathcal{H}_a^{(S)}, \tag{2}$$

其中，$\odot$表示沿嵌入维度进行广播乘法，$\mathbf{w} \in \mathbb{R}^{L_a \times 1}$ 中的每个标量都会缩放 $\mathcal{H}_a^{(T)}$和 $\mathcal{H}_a^{(S)}$中对应音频词元的所有嵌入维度。

计算线性 CKA 的第一步是对加权嵌入进行中心化，即沿词元序列维度减去每个特征的均值。此步骤得到 $\hat{} H_T$ 和 $\hat{} H_S$，分别对应教师模型和学生模型的中心化嵌入。获得中心化嵌入 $\hat{} H_T\hat{}$和 H_S 后，我们使用希尔伯特 - 施密特独立性准则 (HSIC)[22]，通过比较协方差来衡量两个特征之间的相似度。随后，教师和学生的音频嵌入之间的 AwCKA 定义如下：

$$\begin{aligned}
&\text{AwCKA}(\mathcal{H}_a^{(T)}, \mathcal{H}_a^{(S)}, \mathbf{w}) = \text{CKA}(H_T, H_S) \\
&= \frac{\text{HSIC}(H_T, H_S)}{\sqrt{\text{HSIC}(H_T, H_T)} \cdot \sqrt{\text{HSIC}(H_S, H_S)}} \\
&= \frac{\|\hat{H}_T^T \hat{H}_S\|_F^2}{\|\hat{H}_T^T \hat{H}_T\|_F \cdot \|\hat{H}_S^T \hat{H}_S\|_F},
\end{aligned} \tag{3}$$

其中 $\|\cdot\|_F$ 表示弗罗贝尼乌斯范数。可以证明，AwCKA($\mathcal{H}_a^{(T)}$, $\mathcal{H}_a^{(S)}$, $\mathbf{w}$) $\in [0, 1]$ [17]，其值越高，表示两个特征表示之间的相似度越高。

最后，音频投影器输出的蒸馏损失定义如下：

$$\mathcal{L}_{\text{DP}} = 1 - \text{AwCKA}(\mathcal{H}_a^{(T)}, \mathcal{H}_a^{(S)}, \mathbf{w}). \tag{4}$$

2.3. Logits 级蒸馏

如第 2.1 节所述，对于一个音频样本，大语言模型会输出 Logits $Z = [\, Z_a \in \mathbb{R}^{L_a \times V},\ Z_p \in \mathbb{R}^{L_p \times V},\ Z_r \in \mathbb{R}^{L_r \times V}\,]$，其中 L_a、L_p 和 L_r 分别表示音频、提示和响应词元的序列长度，V 表示词汇表大小，总序列长度为 $L = L_a + L_p + L_r$。我们将学生模型和教师模型的 Logits 分别记为 Z_S 和 Z_T。值得注意的是，Logits

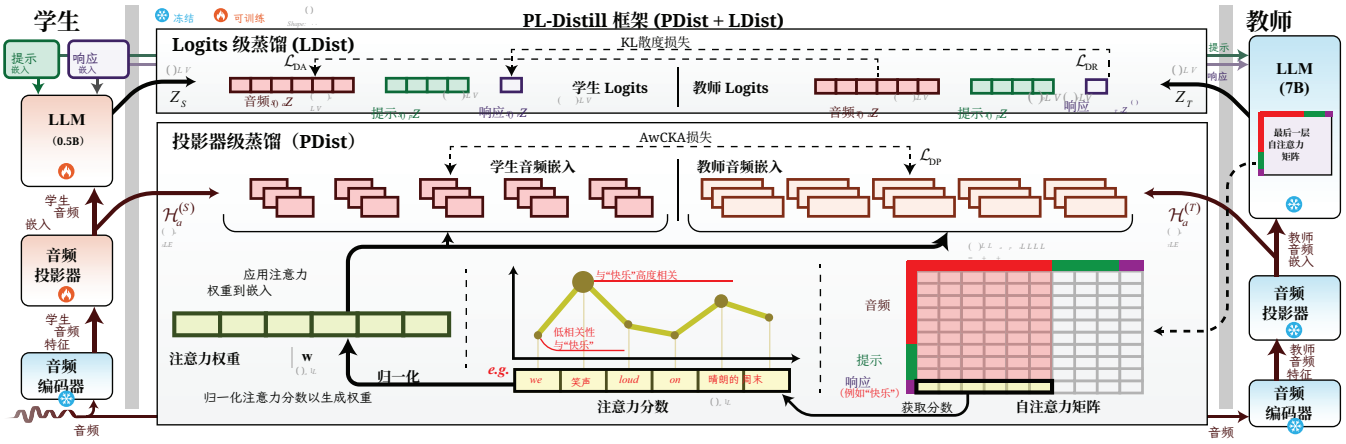


图 1. 我们的 PL- Distill 框架概览，该框架包含投影器级蒸馏（PDist）和 Logits 级蒸馏（LDis t).

教师模型和学生模型的 Logits 具有相同的序列长度和词汇表大小。

在 Logits 级蒸馏（LDist）中，遵循近期关于大语言模型蒸馏的研究，我们将提示词元排除在蒸馏损失之外，因为学生模型无需预测这些词元。对于教师和学生模型的音频 logit（分别记为 $Z_a^{(T)}$和$Z_a^{(S)}$），我们采用温度缩放 softmax，得到相应的概率分布 $P_a^{(T)}$和 $P_a^{(S)}$。随后，我们计算这些分布之间的 Kullback-Leibler（KL）散度，并将其作为蒸馏目标，损失定义为 $\mathcal{L}_{\text{DA}}$。计算过程如下：

$$\mathcal{L}_{\text{DA}} = \text{KL}(P_a^{(T)} \parallel P_a^{(S)}) = \sum_{k=1}^{V} P_a^{(T)}(k) \log \left(\frac{P_a^{(T)}(k)}{P_a^{(S)}(k)} \right),$$

with

$$P_a(k) = \frac{\exp(Z_a(k)/t)}{\sum_{j=1}^{V} \exp(Z_a(j)/t)}. \quad (5)$$

其中， V是大语言模型的词汇表大小，而 $Z_a(k)$ 和 $P_a(k)$ 分别表示词汇表中第 k 个词元对应的 logit 值和概率值。温度 t 控制分布的平滑程度：较高的 t 会产生更平滑的分布，从而揭示更多信息；较低的 t 则会产生更尖锐的分布，突出概率最高的词元 [9]。

与音频 logits 类似，对于响应 logits，我们分别为教师模型和学生模型定义$Z_r^{(T)}$和 $Z_r^{(S)}$，其中$P_r^{(T)}$和 $P_r^{(S)}$表示它们对应的概率分布。记为 $\mathcal{L}_{\text{DR}}$的损失函数可表示为：

$$\mathcal{L}_{\text{DR}} = \text{KL}(P_r^{(T)} \parallel P_r^{(S)}) = \sum_{k=1}^{V} P_r^{(T)}(k) \log \left(\frac{P_r^{(T)}(k)}{P_r^{(S)}(k)} \right). \quad (6)$$

在知识蒸馏中，学生模型不仅从教师模型中蒸馏知识，还会从真实情感标签 [9]中学习。交叉熵损失记为 $\mathcal{L}_{\text{CE}}$，用于衡量学生响应 logits 的概率分布与真实标签之间的差异。 $\mathcal{L}_{\text{CE}}$ 定义如下：

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{L_r} \sum_{k=1}^{V} y_i(k) \log \left(P_r^{(S)}(i,k) \right). \quad (7)$$

其中， L_r 表示响应词元序列的长度。 $P_r^{(S)}(i,k)$表示学生模型预测的第 i个响应词元为词汇表中的第 k个词元的概率。 $y_i(k)$表示第 i个词元真实标签的指示函数，其中，当第i个真实标签词元对应词汇表中的第 k个词元时， $y_i(k) = 1$；否则， $y_i(k) = 0$。

总目标函数为：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \alpha \mathcal{L}_{\text{DP}} + \beta \mathcal{L}_{\text{DA}} + \gamma \mathcal{L}_{\text{DR}}, \quad (8)$$

其中， $\mathcal{L}_{\text{DP}}$ 来自 PDist， $\mathcal{L}_{\text{DA}}$ 和 $\mathcal{L}_{\text{DR}}$ 来自 LDist，而 $\mathcal{L}_{\text{CE}}$ 来自真实情感标签。

3. 实验结果

3.1. 实验设置

在数据集方面，我们采用三个广泛使用的语音情感识别数据集进行分析：IEMOCAP [25], ，包含四种情绪的 5,531 条话语；RAVDESS [26], ，一个视听语料库，包含涵盖八种情绪的 4,800 条语音和歌曲样本；以及 SAVEE [27],，包含七种情绪的 480 条话语，表达方式既有自然的，也有夸张的。

为确保评估公平且具有高度可比性，我们严格遵循 E mobox 基准[23], 的数据划分要求。该基准被公认为迄今范围最广、标准化程度最高的语音情感识别基准。

在训练策略方面，我们冻结学生模型的音频编码器，仅训练学生模型的音频投影器和大语言模型（Qwen2-0.5B），并采用 LoRA 进行高效微调。关于训练参数，损失函数系数 α、 β 和 γ 分别设为 1.0 、 0.8 和 1.0。蒸馏温度 t 设为 2。LoRA 特定参数包括秩 $r = 8$ 和 lora alpha $= 256$。我们采用批次大小为 1、梯度累积步数为 16 的设置，训练五轮，并选取表现最佳的模型进行最终分析。

3.2. 基线与指标

为建立基线并评估训练策略，我们比较了 Emobox 基准中表现最佳的三种预训练语音模型：WavLM large [4], data2vec 2.0 large [24], 和 Whisper large v3 [5]。此外，我们引入一个未经额外训练或微调的教师模型（Qwen2- Audio）作为参考。除我们的蒸馏框架（PL-Distill）外，我们还采用多种策略训练学生模型，以进行比较。这些策略包括监督微调（SFT）、广泛使用的大语言模型蒸馏技术（如前向 KL 和反向KL散度 [11], ），以及

表 1. 各模型在 IEMOCAP、RAVDESS 和 SAVEE 数据集上的主要性能指标对比。预训练模型的结果引自 Emobox 基准 [23]. 对比基线为前向 KL 方法。预训练模型中的当前最优水平结果以粗体标出，所有模型中的总体最佳结果以红色标出。

Type	模型	IEMOCAP			RAVDESS			SAVEE		
		非加权准确率(%)	加权准确率(%)	F1(%)	非加权准确率(%)	加权准确率(%)	F1(%)	非加权准确率(%)	加权准确率(%)	F1(%)
预训练模型	WavLM large [4]	69.47	69.07	69.29	72.00	72.22	71.42	75.65	78.25	**78.38**
	data2vec 2.0 large [24]	57.30	56.23	56.70	71.15	71.63	70.94	**75.75**	**78.59**	78.24
	Whisper large v3 [5]	**73.54**	**72.86**	**73.11**	**75.32**	**75.87**	**75.19**	74.07	77.24	75.31
教师模型	Qwen2-Audio [6]	64.33	60.37	61.61	63.67	63.33	60.84	61.43	58.33	53.92
学生模型	SFT	77.59	76.04	76.29	82.67	76.67	74.48	78.10	80.83	76.70
	前向 KL	79.33	76.77	77.22	85.54	87.08	84.60	83.70	85.00	83.07
	反向KL散度 [11]	79.32	78.34	78.76	87.74	86.67	84.03	79.05	81.67	77.70
	LLaVA-KD [14]	81.56	80.28	80.49	89.36	88.75	88.03	85.71	87.50	85.55
	PL-Distill（我们的方法）	**83.91**	**82.12**	**82.62**	**92.58**	**92.08**	**92.23**	**91.43**	**92.50**	**91.36**
比较	我们的方法与当前最优水平对比	↑14.1%	↑12.7%	↑13.0%	↑22.9%	↑21.4%	↑22.7%	↑20.7%	↑17.7%	↑16.6%
	我们的方法与基线	↑5.77%	↑6.97%	↑6.99%	↑8.23%	↑5.74%	↑9.02%	↑10.3%	↑8.82%	↑10.0%

表2. 在 IEMOCAP、RAVDESS 和 SAVEE 数据集上的消融研究，评估 LDist 和 PDist（采用 CKA 或 AwCKA）。性能指标为非加权准确率和加权准确率（例如，IEMOCAP非加权准确率是 IEMOCAP 的非加权准确率）。

方法	IEMOCAP非加权准确率	IEWA	RAUA	RAWA	SAUA	SAWA
LDist	80.70	79.72	88.24	86.25	83.81	85.83
LDist + PDist（CKA）	81.22	81.11	91.72	91.25	87.62	89.17
LDist + PDist（AwCKA）	83.91	**82.12**	**92.58**	**92.08**	**91.43**	**92.50**

先进的多模态大语言模型蒸馏框架 LLaVA-KD。我们使用三项标准指标评估模型性能：非加权准确率（UA），用于衡量正确分类实例的平均比例；加权准确率（WA），即正确预测数占样本总数的比例；以及宏平均F1（F1），即精确率和召回率的调和平均数。

3.3. 主要性能

如表 1 所示，我们的蒸馏框架 PL-Distill 生成了一个拥有 11 亿参数的学生模型，在所有评估指标（非加权准确率、加权准确率、F1）和数据集（IEMOCAP、RAVDESS 和 SAVEE）上，显著优于教师模型（84 亿参数）、当前最先进的预训练模型以及其他主流蒸馏方法。与预训练模型相比，我们的学生模型在所有指标上均有显著提升。在 RAVDESS 数据集上，其非加权准确率、加权准确率和 F1 分别比表现最佳的预训练模型高出 22.9%、21.4% 和 22.7%。即使是仅经过监督微调的学生模型，其表现也能与最佳预训练模型相当。这些结果证实了大语言模型模块在融合声学和语义特征方面的有效性，展现了大语言模型在语音情感识别中的潜力。PL-Distill 的表现也优于前向 KL 和反向KL散度等主流大语言模型蒸馏方法。例如，在 SAVEE 数据集上，与前向 KL 相比，我们的方法在三项指标上分别提升了 10.3%、8.82% 和 10.0%。这是因为传统方法只关注文本模态，忽视了非文本模态。相比之下，我们的策略 PDist 和 LDist 纳入了音频模态，证明了其对提升性能的重要性。

最后，与最先进的多模态大语言模型蒸馏框架 LLaVA-KD 相比，PL-Distill 也取得了更好的结果。与仅关注蒸馏输出 logits 的现有多模态大语言模型框架不同，PL-Distill 还针对投影器进行蒸馏，有效解决了教师与学生的音频嵌入之间的维度不匹配问题。实验结果表明，投影器作为整合不同模态信息的关键组件，在蒸馏中发挥着至关重要的作用。

3.4. 消融研究

表2展示了在三个数据集上，针对PL-Distill框架核心组件进行的消融研究，并采用非加权准确率和加权准确率作为评估指标。结果清楚地表明，我们提出的投影器级蒸馏（PDist）和AwCKA机制均具有显著效果。具体而言，通过比较LDist和LDist + PDist (CKA)可以发现，加入PDist (CKA)后，模型在所有指标和数据集上的表现均显著提升。这进一步凸显了投影器级蒸馏在实现有效模态对齐方面的关键作用，也证明了CKA方法的有效性。进一步比较LDist + PDist (CKA)与LDist + PDist (AwCKA)，则凸显了AwCKA的显著影响。这支持了我们的假设：在语音情感识别任务中，关键情感特征往往集中于特定的音频词元。通过利用大语言模型最后一个自注意力层的注意力分数对这些词元进行动态加权，AwCKA在CKA计算过程中有效地优先处理这些关键时间步，从而持续取得更优的表现。

4. 结论

本文提出了用于语音情感识别（SER）的大音频语言模型知识蒸馏框架 PL-Distill，成功将一个拥有 8.4B 个参数的教师模型蒸馏为一个拥有 1.1B 个参数的学生模型。我们的方法聚焦于投影器级蒸馏（PDist），并提出了注意力加权中心核对齐（AwCKA），使学生模型能够更好地关注情感相关音频词元。此外，AwCKA 还解决了蒸馏过程中维度不匹配的挑战。实验结果表明，PL-Distill 在多个数据集上均取得了出色的性能。除了语音情感识别之外，PL-Distill 对音频语言建模也具有更广泛的意义。它通过在大音频语言模型中实现有效的知识迁移（连接声学与语言模态），为其他音频任务的高效蒸馏提供了助力。由 AwCKA 驱动的任务相关特征保留策略，也为需要选择性知识迁移的多模态大语言模型蒸馏提供了一种通用方法。未来，我们将探索把 PL-Distill 应用于更多任务。

5. 参考文献

[1]Ahlam Hashem、Muhammad Arif 和 Manal Alghamdi，《语音情感识别方法：系统综述》，《Speech Communication》，第 154 卷，第 102974 页，2023年。[2] Zuheng Kang、Junqing Peng、Jianzong Wang 和 Jing Xiao，《Speecheq：基于多尺度统一数据集和多任务学习的语音情感识别》，载于 INTER-SPEECH 论文集，2022年。[3] Yifu Sun、Xulong Zhang、Jianzong Wang、Ning Cheng、Kaiyu Hu 和 Jing Xiao，《基于分段半监督学习的音乐情感识别研究》，载于 INTERSPEECH 论文集，2023年。[4] Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao 等，《WavLM：面向全栈语音处理的大规模自监督预训练》，《IEEE Journal of Selected Topics in Signal Processing》，第 16 卷，第 6 期，第 1505–1518 页，2022年。[5] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever，《通过大规模弱监督实现稳健的语音识别》，载于 ICML 论文集，2023年，第 28492–28518 页。[6] Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、Yichong Leng、Yuanjun Lv、Jinzheng He、Junyang Lin 等，《Qwen2-Audio 技术报告》，arXiv 预印本 arXiv:2407.10759，2024年。[7] Pengcheng Li、Botao Zhao、Zuheng Kang、Junqing Peng、Xiaoyang Qu、Yayun He 和 Jianzong Wang，《Emo-rl：通过基于情感规则的强化学习增强的广义语音情感识别音频语言模型》，载于计算语言学协会 EMNLP 论文集，2025年，第 18744–18754 页。[8]Sreyan Ghosh、Sonal Kumar、Ashish Seth、Chandra Kiran Reddy Evuru、Utkarsh Tyagi、S Sakshi、Oriol Nieto、Ramani Duraiswami 和 Dinesh Manocha，《Gama：具备先进音频理解和复杂推理能力的大音频语言模型》，载于 EMNLP 论文集，2024年，第 6288–6313 页。[9] G Hinton，《神经网络中的知识蒸馏》，载于 NIPS 深度学习研讨会论文集，2014年。[10] Zuheng Kang、Yayun He、Jianzong Wang、Junqing Peng、Xiaoyang Qu 和 Jing Xiao，《面向通用声音分类无数据知识蒸馏的富特征音频模型反演》，载于 ICASSP 论文集，2023年，第 1–5 页。[11]Yuxian Gu、Li Dong、Furu Wei 和 Minlie Huang，《MiniLLM：大语言模型的知识蒸馏》，载于 ICLR 论文集，2024年。

[12] Tianwei Yin、Micha¨el Gharbi、Richard Zhang、Eli Shechtman、Fredo Durand、William T Freeman 和 Taesung Park，《通过分布匹配蒸馏实现单步扩散》，载于 CVPR 论文集，2024年，第 6613–6623 页。

[13] Fangxun Shu、Yue Liao、Lei Zhang、Le Zhuo、Chenning Xu、Guanghao Zhang、Haonan Shi、Long Chan、TaoZhong、Zhelun Yu、Wanggui He、Siming Fu、Haoyuan Li、Si Liu、Hongsheng Li 和 Hao Jiang，“Llava-mod：通过混合专家知识蒸馏使 Llava 变得微小”，载于 ICLR 论文集，2025年。

[14]Yuxuan Cai、Jiangning Zhang、Haoyang He、Xinwei He、Ao Tong、Zhenye Gan、Chengjie Wang、Zhucun Xue、Yong Liu 和 Xiang Bai，“Llava-kd：多模态大语言模型蒸馏框架”，载于 ICCV 论文集，2025年，第239–249页。[15] Noam Shazeer、Azalia Mirhoseini、Krzysztof Maziarz、Andy Davis、Quoc V. Le、Geoffrey E. Hinton 和 Jeff Dean，“规模惊人的大型神经网络：稀疏门控混合专家层”，载于 ICLR 论文集，2017年。[16] Saurav Muralidharan、Sharath Turuvekere Sreenivas、Raviraj Joshi、Marcin Chochowski、Mostofa Patwary、Mohammad Shoeybi、Bryan Catanzaro、Jan Kautz 和 Pavlo Molchanov，“通过剪枝和知识蒸馏构建紧凑型语言模型”，载于 NeurIPS 论文集，2024年。[17] Simon Kornblith、Mohammad Norouzi、Honglak Lee 和 Geoffrey Hinton，“神经网络表征的相似性再探讨”，载于 ICML 论文集，2019年，第3519–3529页。[18] Jianhan Wu、Xiaoyang Qu、Zhangcheng Huang 和 Jianzong Wang，“通过领域特定软提示生成实现联邦域泛化”，载于 ICCV 论文集，2025年10月，第2366–2375页。[19] Thao Nguyen、Maithra Raghu 和 Simon Kornblith，“宽网络和深网络是否学习相同的内容？揭示神经网络表征如何随宽度和深度变化”，载于 ICLR 论文集，2021年。[20] Max Klabunde、Tobias Schumacher、Markus Strohmaier 和 Florian Lemmerich，“神经网络模型的相似性：功能度量与表征度量综述”，ACM Computing Surveys，第57卷，第9期，第1–52页，2025年。[21] Hassan Sajjad、Nadir Durrani、Fahim Dalvi、Firoj Alam、Abdul Khan 和 Jia Xu，“分析 Transformer 语言模型中的编码概念”，载于 NAACL 论文集，2022年，第3082–3101页。[22] Arthur Gretton、Olivier Bousquet、Alex Smola 和 Bernhard Schölkopf，“使用希尔伯特–施密特范数度量统计依赖”，载于 ALT 论文集，2005年，第63–77页。¨Ziyang Ma、Mingjie Chen、Hezhao Zhang、Zhisheng Zheng、Wenxi Chen、Xiquan Li、Jiaxin Ye、Xie Chen 和 Thomas Hain，“Emobox：多语言多语料库语音情感识别工具包与基准”，载于 INTERSPEECH 论文集，2024年。[23] Alexei Baevski、Arun Babu、Wei-Ning Hsu 和 Michael Auli，“利用上下文化目标表征实现高效自监督学习，应用于视觉、语音和语言”，载于 ICML 论文集，2023年，第1416–1429页。[24] Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N Chang、Sungbok Lee 和 Shrikanth S Narayanan，“IEMOCAP：交互式情感双人动作捕捉数据库”，《语言资源与评估》，第42卷，第359页，2008年，第4期，第335页。[25] Steven R Livingstone 和 Frank A Russo，“Ryerson 音频-视觉情感语音与歌曲数据库（RAVDESS）：北美英语中面部与声音表达的动态多模态集合”，《PloS one》，第13卷，e0196391，2018年，第5期。[26] Philip Jackson 和 SJUoSG Haq，“萨里音频-视觉表达情感（SAVEE）数据库”，萨里大学：吉尔福德，英国，2014年。