arXiv:2601.05011v1 [cs.SD] 8 Jan 2026

利用预测熵实现零样本音频语言分类中的自动提示加权

Karim El Khoury[†,1] *Maxime Zanella*[†,1,2] *Tiffanie Godelaine*[†,1] *Christophe De Vleeschouwer*[1] *Benoˆıt Macq*[1]

[1]ICTEAM，UCLouvain，比利时 [2]ILIA，UMons，比利时

摘要

音频语言模型近期通过利用自然语言监督，在没有带标签训练数据的情况下对音频事件进行分类，展现出强大的零样本能力。然而，其性能对文本提示的措辞高度敏感，细微的变化也可能导致准确率大幅波动。以往的研究通过提示学习或提示集成来缓解这一问题。然而，这些策略要么需要标注数据，要么未能考虑到某些提示可能会对性能产生负面影响。在本文中，我们提出一种熵引导的提示加权方法，旨在寻找提示贡献的稳健组合，以最大化预测置信度。为此，我们构建了一个定制的目标函数，通过最小化预测熵来生成新的提示权重，并利用低熵作为高置信度的代理指标。我们的方法可应用于单个样本或一批音频样本，无需额外标签，且计算开销可以忽略不计。在涵盖环境声音、城市声音和人声声音的五个音频分类数据集上进行的实验表明，在零样本设置下，与经典的提示集成方法相比，该方法取得了持续的提升，整个基准测试中的准确率提升幅度大了5倍。

索引词— 提示加权、熵最小化、零样本分类、音频语言模型

1. 引言

不同于从预定义标签中学习的方法，音频-语言模型（音频语言模型）利用自然语言作为监督信号，更适合描述复杂的真实世界音频录音。这一进展得益于自监督对比预训练方法的发展，以及大规模文本音频数据集的出现 [1–3]。在这些大规模多模态数据集上进行对比预训练，是获得能够在各种音频分类任务中实现良好泛化的模型的强大框架。这种设计使人们能够直接根据类别的文本描述构建音频分类器，而无需进行额外的

额外训练。受 CLIP [4], CLAP [5, 6] 等视觉-语言模型的进展启发，CLAP 是最早研究其在音频领域应用的工作之一 [3, 7–9]。

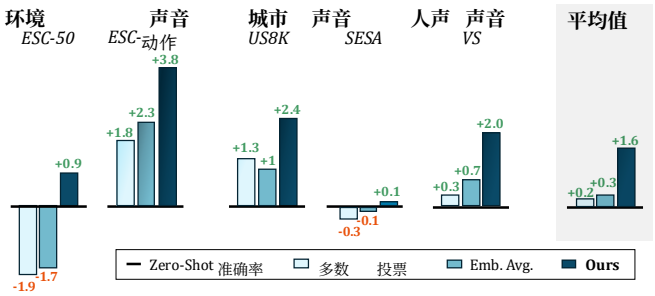

图 1：零样本预测分类准确率提升概览。我们的方法应用于每个数据集的全部数据，相较于多数投票和嵌入平均集成方法，均显示出稳定的改进。

尽管音频语言模型具备可靠的零样本分类准确率 [5], ，但为给定类别设计合适的提示词颇具挑战，甚至会影响最终性能 [10]。事实上，提示词稍有变化，就可能导致预测结果大相径庭，从而使模型性能变得不可靠。缓解这一问题的一种方法是采用少样本提示学习 [10]。然而，这种方法的计算开销可能很大，而且需要标注样本。另一种方法是利用大型语言模型扩充提示词 [11, 12]。尽管这是一个有趣的选择，但它与我们的研究方向互为补充，因为这类方法可以从优化提示加权中获益，进而提升整体性能。第三种方法是使用模板集 [5–7, 13]（表 1 展示了音频语言文献中常见的模板），以降低模型对单个提示词具体措辞的敏感性。由于难以从所有模板中选出合适的提示词，人们通常会在各个提示词预测的基础上采用多数投票，得出最终预测。另一种方法则是简单地对每个类别的各个提示词嵌入取平均值，再使用这些平均嵌入进行分类。

虽然对文本提示取平均值很方便，但显然有些提示的表现远逊于其他提示（见图2）。然而，在纯零样本设置下，要识别表现不佳的提示并非易事。具体而言，我们假设，对文本提示取平均值时，恰当的加权因子应能产生置信度较高的预测。因此，我们提出一种熵最小化目标，旨在找到一个加权向量 β，以构造提示的加权组合，从而得到熵较低的预测。

[†]作者对本工作作出了同等贡献。

致谢 – M.Z. 获得 ARIAC（瓦隆大区资助项目编号：2010235）的资助。T.G. 获得 MedReSyst（欧盟-瓦隆2021—2027计划）的资助。

贡献。我们提出了一个有理论依据的优化框架，将提示集成视为寻找能够最大化预测置信度的权重的问题。具体而言，我们在纯零样本设置下，通过最小化最终预测的熵来实现这一目标。我们的方法能够处理表现不佳的提示，克服传统集成方法的局限，并持续提升准确率。图1总结了我们的结果，展示了该方法在五个音频分类基准上相较于基线集成方法的性能提升。

2. 方法

我们提出一种迭代算法，通过最小化目标函数来求解提示权重向量 β，该目标函数旨在产生低熵（高置信度）的预测。我们通过正则化项对其进行约束，以防止其偏离初始零样本估计，并鼓励权重分布平滑。

2.1. 问题表述

我们首先使用 ALM 音频编码器对每个音频样本 $i \in \{1, \cdots, N_S\}$ 进行编码，得到特征向量 $\mathbf{f}_i \in \mathbb{R}^{1\times d}$。我们获取一组文本嵌入 $\mathbf{t}_{jk} \in \mathbb{R}^{1\times d}$，其中 $\mathbf{t}_{jk}$ 是通过将模板 $j \in \{1, \cdots, N_T\}$ 与类别 $k \in \{1, \cdots, N_C\}$（例如，这是一个 {类别 k 的声音}）组合后，输入 ALM 文本编码器得到的。然后，我们将样本 i 在模板 j 和类别 k 下的 logit l_{ijk} 计算为余弦相似度*:

$$l_{ijk} = \mathbf{f}_i \mathbf{t}_{jk}^T \tag{1}$$

接下来，我们将样本 i在类别 k 下的平均 logit l_{ik} 计算为所有模板j对应 logit 的加权平均。公式如下:

$$\bar{l}_{ik} = \sum_{j=1}^{N_T} \beta_j l_{ijk} \tag{2}$$

其中，$\beta \in \mathbb{R}^{1\times N_T}$ 是加权向量。在均匀情况下，$\beta = \frac{1}{N_T}\mathbf{1}_{N_T}$。应用 softmax 函数后，我们得到样本 i 属于类别 k的预测 p_{ik} :

$$p_{ik} = \frac{\exp(\bar{l}_{ik}/\tau)}{\sum_{l=1}^{N_C}\exp(\bar{l}_{il}/\tau)} \tag{3}$$

其中，τ 是温度缩放超参数，用于控制预测的平滑度。

当 β 均匀时，每个模板对最终 logit 的贡献相同。然而，我们认为这种方法并非最优，特别是在不同模板的分类准确率差异很大时（如图2所示）。因此，我们旨在找到每个模板的最佳贡献，使模型在整个批次的数据样本上输出置信度高（即熵 p_i 低）的预测向量。

> 我们将优化目标表述为寻找一个加权向量 β，
> 使其平均而言能够产生置信度较高的预测。

*注意，所有向量都归一化到单位超球面，因此余弦相似度可简化为点积。

2.2. 目标变量

设变量 $\beta \in \mathbb{R}^{1\times N_T}$ 表示每个不同文本提示模板对应的权重，共有 N_T 个。该向量受约束，必须位于概率单纯形上，并满足以下条件：$\sum_{j=1}^{N_T}\beta_j = 1$和 $\beta_j \geq 0 \quad \forall j$。

2.3. 目标函数

目标是最小化由三项组成的目标函数 $\mathcal{L}(\beta)$：

$$\mathcal{L}(\beta) = \frac{1}{N_S}\sum_{i=1}^{N_S}\Big(\underbrace{H(p_i)}_{\substack{(i)\\ \text{Prediction confidence}}} + \underbrace{\lambda_{zs}H(p_i,\hat{p}_i)}_{\substack{(ii)\\ \text{Zero-shot regularization}}}\Big) - \underbrace{\lambda_\beta H(\beta)}_{\substack{(iii)\\ \text{Entropy regularization}}} \tag{4}$$

(i) 预测置信度。 该项通过最小化每个样本 i的预测向量 p_i 的熵，鼓励所有 N_S 样本都得到置信的预测。

(ii) 零样本正则化。 该项旨在防止预测向量 p_i 偏离初始预测 $\hat{}p_i$过多。通常，我们可以通过单个模板*This is a sound of* {类别} 得到 $\hat{}p_i$，并将其称为零样本预测。

(iii) 熵正则化。 该项是针对共享权重向量 β的基于熵的正则项，在文献中常用 [14–17], , 用于鼓励更平滑、非稀疏的权重分布。由于 β 对所有样本都相同，该项位于求和符号之外。此外，凸惩罚项 $-\lambda_\beta H(\beta)$ 充当熵障碍，从而隐式地强制满足 $\beta_j \geq 0$。

2.4. 优化过程

优化过程的伪代码如算法1所示。需要注意的是，所有嵌入 f和 t均仅在优化开始前计算一次。在初始化时，提示模板权重采用均匀分布进行初始化，并计算零样本预测向量 $\hat{}p$ 。权重 β 根据通过最小化目标函数推导出的迭代不动点更新规则进行更新，直至收敛（即求解 $\frac{\partial\mathcal{L}}{\partial\beta_j} = 0$）：

—

$$\beta_j^{(t)} = \frac{\exp(R_j^{(t)}/\lambda_\beta)}{\sum_{l=1}^{N_T}\exp(R_l^{(t)}/\lambda_\beta)} \tag{5}$$

其中，$R_j^{(t)}$ 可理解为每个独立样本 i所作贡献之和:

$$R_j^{(t)} = \sum_{i=1}^{N_S}\Bigg[\sum_{k=1}^{N_C} p_{ik}\Big(\ln(p_{ik}) + H(p_i)\Big)l_{ijk} + \lambda_{zs}\sum_{k=1}^{N_C} p_{ik}\Big(\ln(\hat{p}_{ik}) - E_{p_i}[\ln(\hat{p}_i)]\Big)l_{ijk}\Bigg] \tag{6}$$

该贡献可分解为两项。第一项与 p_i的熵相关，第二项则评估与零样本预测的相似度，其中 λ_{zs}用于平衡两项的贡献。

算法1： 优化过程

输入： $\mathbf{f}$，$\mathbf{t}$，τ

$\varepsilon \leftarrow 1 \times 10^{-6}$；　▷ 停止阈值

初始化 $\beta^{(0)} = \frac{1}{N_T}\mathbf{1}_{N_T}$；　▷ 均匀权重

计算 $\hat{}p_i \quad \forall i$；　▷ 零样本　预测

当 $\|\beta^{(t)} - \beta^{(t-1)}\|_2 \geq \varepsilon$ **时**

　$t = t + 1$;

　计算 $p_i \quad \forall i$；　▷ See Eq. (3)

　计算 $R_j^{(t)} \quad \forall j$；　▷ See Eq. (6)

　更新 $\beta^{(t)}$；　▷ See Eq. (5)

返回 β；　▷ 优化权重

3. 实验设置

3.1. 评估数据集

我们在涵盖三种不同任务的五个基准音频分类数据集上评估了我们的方法：环境声音、城市声音和人声分类。在环境声音分类方面，我们使用 ESC-50 [18]，，该数据集包含 $2k$ 个音频片段，每段时长为 5 秒，涵盖 50 个类别；还使用 ESC-Actions [18]，，这是前一个数据集的一个子集，包含 400 个音频片段，涵盖 10 个与动作相关的类别。在城市声音分类方面，我们使用 US8K[19]，，该数据集包含 $\sim 8k$ 个音频片段，每段时长为 ~ 4 秒，涵盖 10 个类别；还使用 SESA [20]，，该数据集包含 ~ 600 个音频片段，每段时长为 ~ 30秒，涵盖 4 个监控相关类别。在人声分类方面，我们使用 VS [21]，，该数据集包含 $\sim 21k$个音频片段，分为 6 个类别，每段片段均为 5 秒的录音。对于这五个数据集，我们将整个数据集作为单一测试集，因为我们采用的是零样本设置。

3.2. 文本提示

我们使用音频-语言模型文献中常用的 $N_T = 35$ 手工设计的提示模板 [5–7, 12]构建模板集。这些模板包含多种句式结构，详见表 1。所有五个评估数据集均使用同一组模板。五个基准数据集上的准确率均呈现出较高的变异性（见图 2），凸显了这组模板对于提示集成方法的重要性。

3.3. 实现细节

我们使用原始 CLAP-2022 [5] 音频-语言模型生成音频嵌入$\mathbf{f}$和所有文本嵌入$\mathbf{t}$。根据模型规范，我们将基础模型对数几率缩放设为33.3。关于本方法的超参数，我们在所有实验中将 λ_β 固定为 0.01。至于 λ_{zs}，实验 1（单样本 β）中设为 100，实验2（数据集 β）中设为 0.1。我们将在第 4.2 节进一步讨论这一选择。

3.4. 基线方法

我们实现了六种不同的基线提示集成方法，以便将我们的方法与以下方法进行比较：

(i) 多数投票 指的是，对于每个样本，从所有提示模板中选出预测频率最高的类别。所有提示模板的投票权重相同均等。

表 1：我们实验中使用的提示模板。类别名称应替换 {}。

陈述式（这是……）： 这是一种声音 {}，这是一段音频 {}，这是一段 {}的音频片段，这是一段 {}的声音片段，这是一段音频音轨 {}，声音轨道 {}，这是一个 {}的示例，这是 {}
描述性（A/An……）： 声音： {}，音频： {}，录音： {}，声音录音： {}，音频录音： {}，声音片段： {}，音频片段： {}，音轨： {}，声音轨道： {}，一段声音片段： {}，音频片段： {}
祈使性（听/聆听……）： 聆听 {}，聆听声音 {}，聆听音频： {}，聆听录音： {}，聆听声音 {}的录音，聆听 {}的音频录音，听一听 {}，聆听 {}的音频
直接名词： 声音： {}，音频： {}，录音： {}，声音 {}的录音，{}的音频录音，{}的音频片段
其他： 我能听到 {}，{}

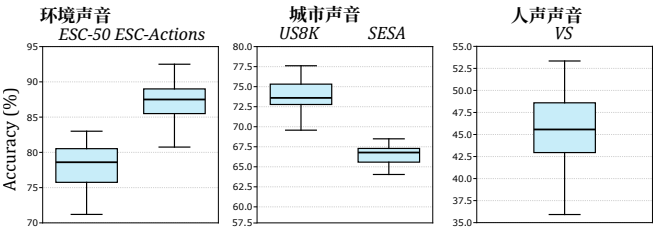

图 2：五个音频数据集中，35 个手工设计的提示模板所获得的准确率分布。结果表明，所有数据集的提示选择都存在显著差异。

（*ii*）熵加权的多数投票则不同，它根据每个模板预测向量熵的倒数为其投票赋予权重。因此，置信度更高的模板预测对最终分类的影响更大。

(iii) 剪枝多数投票 是指在对剩余提示模板进行多数投票之前，移除熵最高的 50% 提示模板。

(iv) 平均文本嵌入 是指为每个提示模板分别计算嵌入，然后针对每个样本计算这些嵌入的归一化总和（即 β 服从均匀分布）。所得的平均文本嵌入用于分类。

(v) 熵加权平均文本嵌入通过在归一化求和时，以对应预测向量熵的倒数为权重，对每个文本嵌入进行加权，从而对这一过程进行调整。因此，由置信度较高的模板生成的文本嵌入，对最终用于分类的嵌入贡献更大。

(vi) 剪枝平均文本嵌入 通过移除熵最高的 50% 提示模板，然后为剩余的每个提示模板计算单独的文本嵌入，最后对这些嵌入求和并归一化，从而得到每个样本对应的结果。所得的平均文本嵌入用于分类。

表2：我们所提方法在五个音频分类数据集上的分类准确率。我们报告三种不同的设置：单样本 β（即 $N_S = 1$）、数据集 β（即 N_S 等于数据集大小），以及带剪枝的数据集 β（更多详情见第4.3节），并与六种基线集成方法进行比较（见第3.4节）。

方法	环境声音 *ESC-50*	*ESC-Actions*	城市声音 *US8K*	*SESA*	人声声音 *VS*	平均值
零样本预测						
“这是一种 {类别}的声音”	82.6	87.7	75.0	66.7	46.9	71.8
基线集成方法						
多数投票	80.7	89.5	76.3	66.4	47.2	72.0
熵加权多数投票	81.2	90.3	75.1	66.6	47.4	72.1
剪枝多数投票	82.9	90.3	75.1	67.1	46.0	72.3
平均文本嵌入	80.5	90.0	76.0	66.6	47.6	72.1
熵加权平均文本嵌入	80.6	90.0	76.2	66.6	47.6	72.2
剪枝平均文本嵌入	82.7	90.3	75.1	67.1	46.6	72.3
所提方法						
单样本 β	82.9	90.0	74.9	**67.2**	47.7	72.5
数据集 β	83.5	90.5	77.3	66.8	47.9	73.2
数据集 β，带剪枝	**83.5**	**91.5**	**77.4**	66.8	**48.9**	**73.6**

4. 实验

4.1. 实验 1：单样本 β

在本实验中，我们考虑每次只能访问一个测试样本 ($N_S = 1$) 的情形，因此为每个样本计算一个单独的加权向量 β。结果如表2所示。我们观察到令人鼓舞的提升：与零样本设置相比，平均增益最高可达 0.7%，表现优于所有基线集成方法。尤其是在 ESC-50 上，我们比平均文本嵌入基线高出 2.4%。这凸显了我们仅凭一个单一样本即可找到优化后的 β 加权向量的能力，优于对文本嵌入进行传统的平均加权。

4.2. 实验2：数据集 β

在本实验中，我们可以访问完整的评估数据集，并为所有样本计算一个通用加权向量 β。我们将 λ_{zs} 的值降至 0.1，以减少对零样本正则化的需求，因为我们使用更多样本计算 β。表3展示了针对 λ_{zs} 参数的消融研究，进一步说明了我们选择该参数的理由。如表2所示，与单样本 β 优化相比，我们平均提升了 0.7%；与所有基线方法相比，平均提升了0.9%。

4.3. 实验3：数据集 β 剪枝

在本实验中，我们对加权向量 β 执行多个剪枝周期。在每个周期中，我们先优化 β，然后对其进行剪枝。剪枝后的 β 向量用于初始化下一个周期。我们进行4个周期，每个周期的剪枝比例为15%，最终总体剪枝比例达到 ~ 50%。此外，我们还进行了消融研究（见表4），以进一步验证我们的选择。如表2所示，与数据集级 β 优化相比，我们平均又提升了0.4%，使所有基线集成方法的累计平均提升达到1.4%。值得注意的是，我们在5个数据集中的4个上取得了最佳表现，其中 ESC-Actions 相较于其零样本预测最高提升了3.8%。请注意，即使经过多个剪枝周期，所提方法的运行时间仍与基线方法相当，为0.2秒（见表5）；与编码音频和文本特征所需的时间相比，这一耗时仍可忽略不计。

表3：不同 λ_{zs} 取值下所有数据集的平均值准确率，突显了在单样本优化情形下，需要进一步正则化以趋向零样本预测。蓝色标出所选项。

	λ_{zs} *0.01*	*0.1*	*1*	*10*	*100*	*1000*
单样本 β	72.18	72.22	72.38	72.39	**72.48**	72.22
数据集 β	72.79	**73.19**	71.33	70.84	70.62	70.61

表4：每轮不同剪枝比例及总轮数下所有数据集的平均值准确率。蓝色标出所选项。

每轮剪枝比例	总轮数 *1*	*2*	*3*	*4*	*5*
5	73.26	72.92	73.13	73.33	73.30
10	72.96	73.31	73.54	73.42	73.45
15	73.22	73.35	73.52	**73.60**	73.56
20	73.37	73.45	73.43	73.34	73.20
25	73.46	73.42	73.38	72.21	72.12

表5：与特征提取和基线方法相比，所提方法的运行时间。在使用包含 35 个模板的设置下，采用 VS数据集（~ 21k 个音频样本分为六个类别），并在 24 GB NVIDIA GeForce RTX 4090 GPU 上进行评估。

	特征编码	基线方法	所提方法
运行时间	~ 2分钟	~ 0s	~ 0.2s

5. 结论

我们提出一种用于零样本音频-语言分类的熵引导提示加权算法，通过将低熵预测作为高置信度的代理，策略性地融合来自多个模板的嵌入。我们证明，该方法在五个基准音频分类数据集上始终优于基线集成方法。所提出的方法提供了一种实用方案，可集成到现有音频-语言模型的零样本分类流程中，从而在遇到新数据集时无需手动设计提示词。

参考文献

[1] Chris Dongjoo Kim、Byeongchang Kim、Hyunmin Lee 和 Gunhee Kim，“Audiocaps：为真实场景中的音频生成说明文字”，载于《*2019* 年计算语言学协会北美分会会议论文集：人类语言技术》，第 *1* 卷（长篇和短篇论文），2019 年，第 119–132 页。

[2] Konstantinos Drossos、Samuel Lipping 和 Tuomas Vir- tanen，“Clotho：一个音频描述数据集”，载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2020年，第736–740页。

[3] Yusong Wu、Ke Chen、Tianyu Zhang、Yuchen Hui、Tay- lor Berg-Kirkpatrick 和 Shlomo Dubnov，“通过特征融合和关键词到描述增强进行大规模对比式语言-音频预训练”，载于*ICASSP 2023-2023 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2023年，第1–5页。[4] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sas- try、Amanda Askell、Pamela Mishkin、Jack Clark 等，“从自然语言监督中学习可迁移的视觉模型”，载于国际机器学习会议。ICML，2021年，第8748–8763页。[5] Benjamin Elizalde、Soham Deshmukh、Mahmoud Al Is- mail 和 Huaming Wang，“CLAP：从自然语言监督中学习音频概念”，载于*ICASSP 2023 - 2023 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，2023年，第1–5页。[6]Benjamin Elizalde、Soham Deshmukh 和 Huaming Wang，“面向通用音频表征的自然语言监督”，载于*ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2024年，第336–340页。[7] Soham Deshmukh、Benjamin Elizalde、Rita Singh 和 Huaming Wang，“Pengi：面向音频任务的音频语言模型”，神经信息处理系统进展，第36卷，第18090–18108页，2023年。[8] Ching-Feng Yeh、Po-Yao Huang、Vasu Sharma、Shang- Wen Li 和 Gargi Gosh，“FLAP：快速语言-音频预训练”，载于*2023* 年 *IEEE*自动语音识别与理解研讨会 *(ASRU)*。IEEE，2023年，第1–8页。[9] Xuenan Xu、Zhiling Zhang、Zelin Zhou、Pingyue Zhang、Zeyu Xie、Mengyue Wu 和 Kenny Q Zhu，“BLAT：基于 AudioSet 标签引导的合成数据引导语言-音频预训练”，载于第*31*届 *ACM*多媒体国际会议论文集，2023年，第2756–2764页。

[10] Asif Hanif、MahaTufail Agro、Mohammad AreebQazi 和 Hanan Aldarmaki，“Palm：音频语言模型的少样本提示学习”，*arXiv* 预印本 *arXiv:2409.19806*，2024年。

[11] Nishit Anand、Ashish Seth、Ramani Duraiswami 和 Di- nesh Manocha，“Tspe：通过任务特定提示集成提升零样本音频分类”，发表于*2025 IEEE*声学、语音与信号处理国际会议研讨会（*ICASSPW*），2025年，第1–5页。[12] Michel Olvera、Paraskevas Stamatiadis 和 Slim Essid，“声音描述：探索提示模板和类别描述以增强零样本音频分类”，*arXiv* 预印本 *arXiv:2409.13676*，2024年。[13] James Urquhart Allingham、Jie Ren、Michael W Dusen- berry、Xiuye Gu、Yin Cui、Dustin Tran、Jeremiah Zhe Liu 和 Balaji Lakshminarayanan，“一种简单的零样本提示加权技术，可改进文本-图像模型中的提示集成”，发表于国际机器学习会议。ICML，2023年，第547–568页。[14] S´egol`ene Martin、Yunshi Huang、Fereshteh Shakeri、Jean- Christophe Pesquet 和 Ismail Ben Ayed，“传导式零样本与少样本 CLIP”，发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第28816–28826页。[15] Maxime Zanella、Benoˆıt G´erin 和 Ismail Ayed，“利用传导增强视觉-语言模型”，神经信息处理系统进展，第37卷，第62223–62256页，2024年。[16]Maxime Zanella、Fereshteh Shakeri、Yunshi Huang、Houda Bahig 和 Ismail Ben Ayed，“增强用于组织病理学分类的视觉-语言模型：一次性预测全部类别”，发表于通用医疗人工智能基础模型国际研讨会。施普林格，2024年，第153–162页。[17] Karim El Khoury、Maxime Zanella、Benoˆıt G´erin、Tiffanie Godelaine、Benoˆıt Macq、Sa¨ıd Mahmoudi、Christophe De Vleeschouwer 和 Ismail Ben Ayed，“增强用于零样本场景分类的遥感视觉-语言模型”，发表于*ICASSP 2025-2025 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2025年，第1–5页。[18] Karol J. Piczak，“ESC：用于环境声音分类的数据集”，发表于第*23* 届*ACM*多媒体年会论文集，第1015–1018页，ACM出版社。[19]Justin Salamon、Christopher Jacoby 和 Juan Pablo Bello，“用于城市声音研究的数据集与分类体系”，发表于第*22*届*ACM*多媒体国际会议论文集，2014年，第1041–1044页。[20] Tito Spadini，“用于监控应用的声音事件”，2019年10月，2019年10月。[21] Yuan Gong、Jin Yu 和 Ja mes Glass，“VocalSound：用于提升人类发声识别的数据集”，发表于*ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022年，第151–155页。