arXiv:2303.02344v1 [cs.CV] 4 Mar 2023

利用伪视觉标签改进视听视频解析

周金星、郭丹、钟亦然和王孟，*IEEE* 会士

摘要—**视听视频解析旨在预测各模态下视频片段中发生的事件。该任务通常采用弱监督方式，即仅提供视频事件标签，即，而标签所属的模态及其时间戳均未知。由于缺乏密集标注的标签，近期研究尝试利用伪标签来丰富监督信息。一种常用策略是根据已知事件标签，为各模态生成伪标签。然而，这些标签仍仅限于视频级，事件时间戳的时间边界依然没有标注。本文提出一种新的伪标签生成策略，利用从开放世界中学习到的先验知识，为每个视频片段显式分配标签。具体而言，我们利用 CLIP 模型，基于视觉模态估计每个视频片段中的事件，从而生成片段级伪标签。我们提出一种新的损失函数，通过考虑标签的类别丰富度和片段丰富度来正则化这些标签。我们采用标签去噪策略，在出现较高的前向二元交叉熵损失时翻转伪标签，以改进伪标签。我们在 LLP 数据集上进行了大量实验，结果表明，借助新提出的损失函数和标签去噪策略，我们的方法能够生成高质量的片段级伪标签。我们的方法在视听视频解析方面达到了最先进水平。**

索引词——**视听视频解析、伪标签、标签去噪、音视频学习。**

I. 引言

音频（AUDIO）和视觉信号大量涌入我们的生活，每种信号都可能承载各种事件。例如，我们经常会看到街上的汽车和行人，同时也能听到汽车喇叭声和人们交谈的声音。人类之所以能够全面地理解场景，很大程度上得益于同时使用听觉和视觉传感器。为了让机器具备这类智能，许多研究从单模态理解的一些基础任务入手，例如如音频事件分类 [1]–[4], 、视频分类 [5]–[8], 和时序动作定位 [9]–[12]。音频事件分类任务侧重于音频模态的识别，而视频分类和动作定位任务侧重于视觉模态。随着研究不断深入，许多工作进一步探索了多模态视听感知 [13],, 催生了声源定位 [14]–[24], 、视听事件定位 [25]–[36], 和视听问答 [37]–[39]等任务。

最近，Tian 等人 [40] 提出了一项新的多模态场景理解任务，即视听视频解析（AVVP）。给定一个可听视频，AVVP任务要求识别音频与视觉轨道中发生了哪些事件，以及这些事件发生在哪些视频片段中。因此，需要预测每种模态中各事件的类别和时间边界。需要注意的是，音频与视觉轨道都可能包含多个不同的事件，且事件通常存在于连续片段中，因此，为每种模态提供片段级事件标签是一项耗时的工作。为此，AVVP在弱监督设置下进行，仅提供视频标签。如图1（a）所示的示例，我们只知道该视频包含语音和吸尘器事件。对于每个事件，模型需要判断其存在于音频模态（音频事件）、视觉模态（视觉事件）还是两种模态中（视听事件），并定位具体的时间片段，如图1（b）所示。值得注意的是，视听事件是音频事件与视觉事件的交集。

在这项工作中，我们强调，视听视频解析任务面临两大主要挑战。**1）视频标签带来的干扰**。如图 1（b）所示，事件 语音同时发生在音频与视觉模态中，而吸尘器则仅存在于视觉模态中。在数据标注时，音频和视觉轨道共享来自视频标签的相同监督，即，{语音，吸尘器}。因此，在模型训练过程中，视觉对象吸尘器会干扰音频事件解析。同样，视频标签也可能干扰视觉事件解析。**2）时序片段区分**。假设我们识别出视觉模态中存在吸尘器事件，在弱监督标签（视频级）的条件下，仍然难以区分哪些片段包含该事件（片段级）。这些挑战使视听视频解析任务成为一个难以处理的多模态多实例学习（MMIL）问题。

在开创性工作中， [40],提出了混合注意力网络（HAN）来编码视听时序特征，随后通过对这些特征进行注意力池化来预测视频

周金星现任职于合肥工业大学（HFUT）教育部大数据知识工程重点实验室、计算机科学与信息工程学院（人工智能学院），以及安徽省智能互联系统实验室（HFUT），地址：中国合肥，邮编230601。电子邮箱：zhoujxhfut@gmail.com。

郭丹和王孟就职于大数据知识工程重点实验室（合肥工业大学），教育部，合肥工业大学计算机科学与信息工程学院（人工智能学院），安徽省智能互联系统实验室（合肥工业大学）以及合肥综合性国家科学中心，地址：中国合肥市230601。电子邮箱：guodan@hfut.edu.cn, eric.mengwang@gmail.com。

钟亦然现供职于中国上海市上海人工智能实验室，邮编 200000。电子邮箱：zhongyiran@gmail.com。

通信作者：郭丹和王孟 .

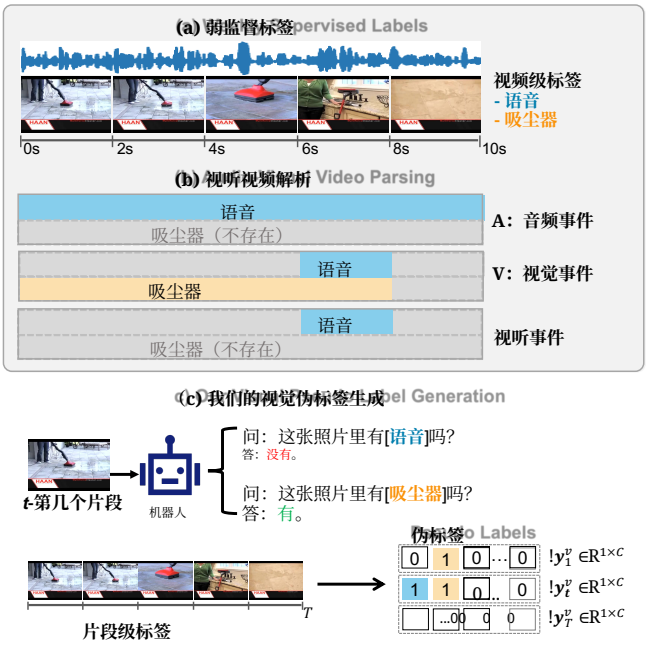


图 1. **弱监督音视频解析（AVVP）任务及我们的伪标签生成过程示意图。** (a) 给定一个视频及其事件标签（“语音”和“吸尘器”），(b) AVVP 需要预测音频事件、视觉事件和视听事件。请注意，“吸尘器”仅存在于视觉轨道中，而“语音”同时存在于音频与视觉轨道中，因此形成视听事件“语音”。(c) 在我们的视觉伪标签生成过程中，我们使用 CLIP 判断每个视频片段中发生了哪些事件。随后，我们获得可靠的片段级视觉伪标签，以简化弱监督视频解析任务。

事件。弱视频标签被用作主要监督。为应对这一挑战，他们提出通过标签平滑 [41] 技术处理视频标签，从而为独立的音频与视觉模态获取伪标签。实验表明，为每种模态生成伪标签能够为监督事件解析带来显著益处 [40]。后续研究分化为两个方向。其中大多数研究致力于设计有效的网络，以隐式聚合多模态特征进行预测 [42]–[47]，，同时使用由 HAN 生成的伪标签 [40]。相比之下，其他新近工作 [48], [49] 则致力于基于 HAN 的基线网络为每种模态生成伪标签。生成的伪标签根据已知的视频标签进行去噪，因此仅限于视频级别，只能表明每种模态中存在哪些事件。要区分事件包含在哪些片段中，仍然十分困难。

为应对上述挑战，我们的工作始于这样一种直觉：能否为每个片段至少针对一种模态显式生成伪标签，从而促进这一 MMIL 任务的完成？这一想法受到以下两点观察的启发：1）音视频事件理解模型有望在片段级标签的引导下取得更好的表现，因为片段级标签能够提供更明确的监督信息；2）在这一多模态任务中，即使只提供某一种模态的一些微小提示，人类也能显著增强视频理解。与音频相比，视觉信号通常在图像中包含更明确的语义，因此即使在片段级别也更容易进行分类。此外，在实践中，很难生成音频模态的片段级伪标签。原因在于，音频信号是连续的，难以沿时间轴进行分离，而且主流的音频事件分类模型通常针对整个视频而非每个片段 [1]–[4]进行类别分类。为此，我们提出了一种**视觉伪标签探索（Visual Pseudo LAbel exploratioN，VPLAN）方法**，旨在为视觉模态生成高质量的片段级伪标签，并进一步助力多模态 AVVP 任务。

为了获得视觉伪标签，一个自然的想法是借助图像分类预训练模型中的现成知识，例如各种分类模型，或在 ImageNet 上预训练的视觉变换器 [50]。给定从 AVVP 任务中的视频样本提取的一幅图像，此类预训练模型可以将图像分类到 ImageNet 的预定义类别中，从而获得帧级或片段级标签。然而，源数据与目标数据的类别词汇表并不一致。因此，直接使用图像分类模型会引入不可忽视的标签噪声。近年来，视觉语言预训练备受关注，它从大规模网络数据中学习视觉图像和文本的高级表征，推动了许多跨模态下游任务的发展 [51]–[53]。作为代表性工作，对比语言-图像预训练（CLIP） [53]开创了零样本图像分类的新范式。给定一幅图像，将其可能的类别名称填入预定义的文本提示中。随后，CLIP 根据编码后的文本与图像特征之间的相似度为各类别打分。CLIP 能够灵活地对开放类别词汇表中的图像进行分类。

受此启发，我们提出一种**伪标签生成（PLG）**模块，利用 CLIP [53] 生成片段级视觉伪标签[1]。图 1 (c) 展示了一个简单示例。给定所有潜在的事件标签，我们让 CLIP 扮演智能机器人的角色，判断给定片段中是否包含相应事件。最终，得分较高的查询事件类别被视为视觉伪标签。此过程可应用于视频帧或视频片段。在本研究中，我们将视频划分为多个视频片段，并探索片段级标签。更多实现细节见第 IV-A 节。生成的伪标签用于为模型训练提供细粒度监督。此外，我们发现，获得的片段级伪标签包含丰富的信息，表明每个片段中发生了多少类事件（类别丰富度），以及某一事件类别出现在多少个片段中（片段丰富度）。如图 1 (b) 所示，视频级标签表明视觉轨道中最多可能有两个事件，即，语音和吸尘器。实际上，只有第四个片段同时包含两个事件，而第一个片段包含一个事件——吸尘器。因此，第四个和第一个片段的类别丰富度分别可记为 1 和 1/2。同样，从事件类别的角度来看，吸尘器出现在整个视频的四个视频片段中（共五个片段），而

[1]除非另有说明，下文中的“伪标签”均指视觉伪标签。

语音只出现在一个（第四个）片段中。因此，吸尘器和语音事件的片段丰富度分别可记为 4/5 和 1/5。模型应能识别类别丰富度和片段丰富度的差异，从而做出正确预测。基于此，我们提出一种新颖的**丰富度感知损失（RL）**，用于使预测中包含的丰富度信息与伪标签中的丰富度信息保持一致。实验验证，生成的伪标签结合 RL 损失能够提升视频解析性能。最后，考虑到 PLG 生成的伪标签无需人工参与，且可能含有噪声，我们进一步提出**伪标签去噪（PLD）**策略来改进伪标签。带有噪声标签的样本通常难以学习，且会产生较大的前向传播损失 [54]–[56]。在我们的研究中，较大的损失可能来自被分配了错误伪标签的难学习数据。PLD 旨在改进这些数据的伪标签。更多实现细节见第 IV-B 节。

我们在公开的 LLP 数据集 [40] 上评估了我们的方法。实验结果证明了主要设计的有效性。我们的方法在视觉事件和视听事件解析方面达到了新的最先进水平，同时在音频事件解析方面也保持了竞争力。据我们所知，我们是首个探索片段级视觉伪标签策略的研究。值得注意的是，PLG 生成并由 PLD 和所提出的 RL 损失进一步优化的伪标签，可以轻松应用于现有的视听视频解析任务方法。

II. 相关工作

视听视频解析（AVVP）。视听视频解析任务需要识别每种模态中发生的事件，并定位事件所在的对应视频片段。Tian等人 [40] 首次提出了这一任务，并设计了一种混合注意力网络，用于聚合模态内和模态间特征。此外，他们采用标签平滑 [41] 策略，以解决单一视频级标签造成的模态标签偏差。一些方法侧重于网络设计。Yu 等人 [45] 提出了一种多模态金字塔注意力网络，其中包含多个用于编码时序特征的金字塔单元。Jiang 等人 [47] 使用两个额外且独立的视觉和音频预测网络，以缓解音频与视觉模态之间的标签干扰。Mo等人 [42] 使用可学习的类别感知令牌，对来自彼此独立的音频与视觉模态的语义进行分组。为克服标签干扰，Wu等人 [48] 交换两个事件相互独立的视频的音频与视觉轨道，以构造用于模型训练的新数据。随后根据重构视频的预测结果生成伪标签。Cheng 等人 [49] 首先估计视频标签的噪声比例，并反转前向损失较大的特定比例标签。尽管这些方法带来了显著提升，但它们只能生成视频级事件标签。与之不同的是，我们旨在直接获取片段级的高质量视觉模态伪标签，进而助力整个视频解析系统的训练。

使用伪标签进行学习。深度神经网络在各种任务中都取得了卓越的表现，这在很大程度上要归功于可用于训练的大量标注数据。伪标签方法旨在利用海量无标注数据进一步提升性能。通常，首先使用标注数据训练一个知识模型，然后用它为无标注数据生成伪标签。最后，使用标注数据和伪标注数据训练一个新模型。这一策略已被证明有益于多种任务，例如图像分类 [57]–[62]、语音识别 [63], [64],以及基于图像的文本识别 [65]。在这项工作中，我们使用预训练大模型作为知识模型，为视听视频解析任务所用的数据生成伪标签，再利用伪标注数据更好地训练解析模型。生成的伪标签可能仍然含有噪声。一些研究探索了在噪声标签下进行模型学习时的训练损失。Hu 等人 [55]提出通过提高干净样本的权重、降低难学习样本的权重来优化网络。在弱监督多标签分类问题中，Kim 等人 [56] 提出纠正那些可能具有较大损失的假阴性标签。受这些工作的启发，我们设计了一种标签去噪策略，以应对生成的片段级伪标签中潜在的噪声。

CLIP 预训练。 这里我们也讨论预训练技术，并详细阐述选择 CLIP 作为生成伪标签的基础大模型的原因。CLIP [53] 使用对比学习技术，在包含 4 亿个图文对的数据集上进行训练。这种大规模预训练使 CLIP 能够学习图像和文本的高效表征，并在零样本图像分类上展现出令人瞩目的性能。其零样本迁移能力为解决许多任务开辟了新途径，并催生了大量研究工作，例如图像描述 [66],视频描述 [67],和语义分割 [68]–[72]。大多数研究选择冻结或微调 CLIP 的图像和文本编码器，以提取用于下游任务 [66]–[68], [73],[74]的高级特征。对于零样本语义分割，一些方法开始使用预训练 CLIP 生成像素级伪标签，这些伪标签无需标注且有助于任务完成 [71], [72]。在这项工作中，我们首次尝试借助 CLIP 的先验知识，以简化具有挑战性的视听视频解析任务。

III. 预备知识

本节将详细阐述 AVVP 任务，并简要介绍基线框架 HAN [40] 。该框架既用于我们的方法，也用于 AVVP 任务中的这些伪标签生成工作 [48], [49]。

任务定义。 AVVP 任务旨在定位视频中音频与视觉轨道对应的视频片段中的事件，并对事件类别进行分类。具体而言，给定一段时长为 T 秒的视频序列 $\{V_t, A_t\}_{t=1}^{T}$，V_t 和 A_t 分别表示第 t 个视频片段中的视觉组件和音频组件。视频 $\boldsymbol{y}^{v\cup a} \in \mathbb{R}^{1\times C} = \{y_c^{v\cup a} | y_c^{v\cup a} \in \{0,1\}, c = 1, 2, ..., C\}$的事件标签中， C是事件类别的总数，上标 $v \cup a$ 表示整个视频的事件标签是音频与视觉模态标签的并集， $y_c^{v\cup a}$的值为 1 表示视频中发生了第

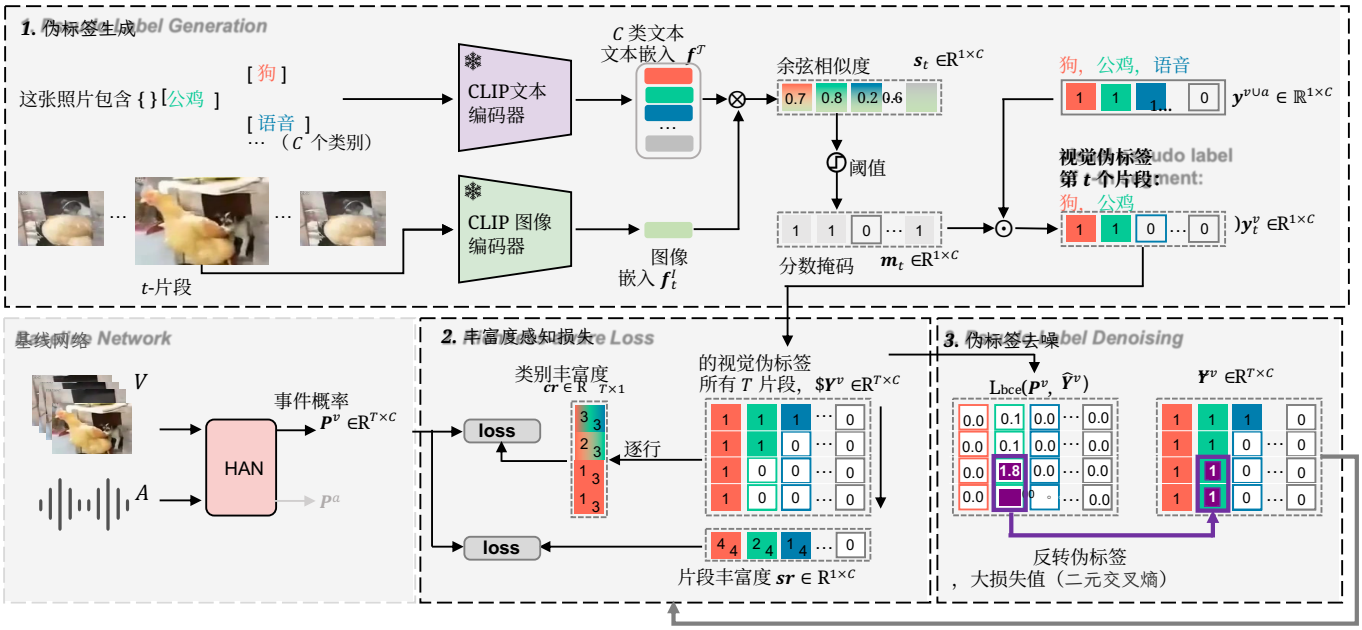


图2。**我们方法的概述**。作为一种标签细化方法，我们旨在生成高质量、细粒度的片段级视频标签。对于骨干网络，可使用任何适用于 AVVP 任务的网络来生成事件预测。这里，我们采用基线 HAN [40]。在我们的方法中，我们设计了一个伪标签生成模块，利用预训练 CLIP [53] 为视觉模态生成片段级伪标签。例如，对于视觉轨道中的第 t 个片段，视频标签“语音”会从视觉模态中滤除，因为 CLIP 计算出的该类别相似度分数较低。值得注意的是，CLIP 仅用于伪标签生成，且其参数被冻结。随后，利用生成的视觉伪标签，我们提出一种丰富度感知损失，作为一种新的全监督方式，帮助模型使视觉预测与伪标签中的类别丰富度和片段丰富度相匹配。最后，我们设计了一种伪标签去噪策略，通过交换 $\mathcal{L}_{\text{bce}}$ 估计中前向损失值较大的位置（即二元交叉熵较大），进一步细化伪标签。经 PLD 细化后的伪标签可进一步用作模型训练的新监督信号。 $\otimes$ 表示矩阵乘法， $\odot$ 表示逐元素乘法。

c 类事件。注意， $\boldsymbol{y}^{v\cup a}$ 是视频级的弱监督标签，训练期间，每个视频片段中各个模态的标签未知。然而，为了进行评估，需要预测每个片段中包含的音频事件、视觉事件和视听事件。我们将视频级视觉与音频事件的概率分别表示为 $\{\{\boldsymbol{p}^v;\ \boldsymbol{p}^a\} \in \mathbb{R}^{1\times C} | p_c^v,\ p_c^a \in [0,1]\}$，$\boldsymbol{p}^{v\cap a} = \boldsymbol{p}^v * \boldsymbol{p}^a$ 用于表示它们的交集。因此，所有视频片段中视觉事件、音频事件和视听事件的概率可分别表示为 $\{\boldsymbol{P}^v;\ \boldsymbol{P}^a;\ \boldsymbol{P}^{v\cap a}\} \in \mathbb{R}^{T\times C}$，这些概率都需要进行预测。

基线框架。基线网络 HAN [40] 使用 Transformer [75] 中的多头注意力（MHA）机制，对音频与视觉模态的模态内特征和跨模态特征进行编码。我们将通过预训练卷积神经网络提取的初始音频和视觉特征分别记为 $\boldsymbol{F}^a$、$\boldsymbol{F}^v \in \mathbb{R}^{T\times d}$，其中 d 为特征维度。HAN 的过程可概括为：

$$\begin{cases} \dot{\boldsymbol{F}}^a = \boldsymbol{F}^a + \texttt{MHA}(\boldsymbol{F}^a, \boldsymbol{F}^a) + \texttt{MHA}(\boldsymbol{F}^a, \boldsymbol{F}^v), \\ \dot{\boldsymbol{F}}^v = \boldsymbol{F}^v + \texttt{MHA}(\boldsymbol{F}^v, \boldsymbol{F}^v) + \texttt{MHA}(\boldsymbol{F}^v, \boldsymbol{F}^a), \end{cases} \tag{1}$$

其中， $\dot{\boldsymbol{F}}^{\boldsymbol{a}}$, $\dot{\boldsymbol{F}}^{\boldsymbol{v}} \in \mathbb{R}^{T\times d}$ 是更新后的音频与视觉特征。通过全连接（FC）层和 Sigmoid函数，可以预测音频与视觉模态的片段级事件概率 $\boldsymbol{P}^a \in \mathbb{R}^{T\times C}$ 和 $\boldsymbol{P}^v \in \mathbb{R}^{T\times C}$。文献 [40] 进一步提出了注意力多实例多标签池化，用于对时序音视频特征进行加权和聚合。随后，我们得到整个视频中音频的最终视频级预测事件、视觉事件及其并集事件，分别记为 $\{\boldsymbol{p}^a; \boldsymbol{p}^v; \boldsymbol{p}^{v\cup a}\} \in \mathbb{R}^{1\times C}$。模型训练的目标函数为：

$$\mathcal{L} = \mathcal{L}_{\text{bce}}(\boldsymbol{p}^{v\cup a}, \boldsymbol{y}^{v\cup a}) + \mathcal{L}_{\text{bce}}(\boldsymbol{p}^a, \overline{\boldsymbol{y}}^a) + \mathcal{L}_{\text{bce}}(\boldsymbol{p}^v, \overline{\boldsymbol{y}}^v), \tag{2}$$

其中， $\mathcal{L}_{\text{bce}}$ 为二元交叉熵损失， $\{\boldsymbol{y}^v, \boldsymbol{y}^a\} \in \mathbb{R}^{1\times C}$是通过对 $\boldsymbol{y}^{v\cup a}$进行标签平滑 [41] 生成的视频级视觉和音频伪标签。

IV. 我们的方法

我们的方法概览如图2所示。我们致力于生成可靠的片段级伪标签，以更好地监督模型进行视听视频解析。对于骨干网络，我们仅采用基线 HAN [40] 来输出预测。我们的方法带来了以下创新。1）我们提出了一种**伪标签生成（PLG）**模块，利用预训练 CLIP 模型从片段级生成可靠的视觉伪标签。2）为利用所获得的伪标签，我们设计了一种**丰富度感知损失（RL）**用于模型训练。该损失有助于对预测进行正则化，使其能够感知伪标签中包含的类别丰富度和片段丰富度。3）我们进一步提出了一种**伪标签去噪（PLD）**策略，通过翻转难学习数据的伪标签来优化生成的伪标签；这些数据的$\mathcal{L}_{\text{bce}}$ 估计前向损失值较高（即二元交叉熵较大）。

A. 伪标签生成

PLG 旨在生成高质量的片段级视觉伪标签，以期缓解视频标签带来的标签干扰，并更好地监督模型区分不同的视频片段。如第 I 节所述，我们选择预训练 CLIP [53] 模型来实现这一目标，因为它具备灵活的零样本分类能力。下面详细介绍伪标签生成过程。

具体而言，每个视频样本被均匀划分为若干个视频片段，并从每个视频片段中采样一帧作为其代表。如图 2-1 所示，对于第 t 个片段中采样的帧 $\mathcal{I}_t$，使用 CLIP图像编码器提取图像特征，记为 $\boldsymbol{f}_t^{\mathcal{I}} \in \mathbb{R}^{1\times d}$。CLIP文本编码器的文本输入采用提示词“A photo of a [CLS]”，其中 [CLS] 可替换为潜在的类别名称。对于视听视频解析任务，我们根据经验将提示词改为更合适的“This photo contains the[CLS]”（CLIP文本编码器中提示词的消融研究见第 V-B 节）。将此提示词中的 [CLS] 替换为各个事件类别名称，并将生成的文本输入 CLIP文本编码器，即可获得所有 C 类 $\boldsymbol{f}^{\mathcal{T}} \in \mathbb{R}^{C\times d}$ 的文本（类别）特征。然后，可以计算图像与事件类别之间的归一化余弦相似度 $\boldsymbol{s}_t \in \mathbb{R}^{1\times C}$，如下所示：

$$\boldsymbol{s}_t = softmax(\frac{\boldsymbol{f}_t^{\mathcal{I}}}{\|\boldsymbol{f}_t^{\mathcal{I}}\|_2} \otimes (\frac{\boldsymbol{f}^{\mathcal{T}}}{\|\boldsymbol{f}^{\mathcal{T}}\|_2})^{\top}), \tag{3}$$

其中，$\otimes$ 表示矩阵乘法，$\top$ 表示矩阵转置。在 $\boldsymbol{s}_t$ 中，较高的相似度分数表明该事件类别更有可能出现在图像中。

此外，我们使用阈值 τ 从 $\boldsymbol{s}_t$ 中选出分数较高的类别，并获得分数掩码 $\boldsymbol{m}_t$。随后，我们通过逐元素乘法 $\odot$ 将分数掩码 $\boldsymbol{m}_t$ 应用于视频级标签$\boldsymbol{y}^{v\cup a}$，以筛除出现在第 t个片段$\hat{}\boldsymbol{y}_t^v \in \mathbb{R}^{1\times C}$的视觉事件。该过程可表示为：

$$\begin{cases} \boldsymbol{m}_t = \mathbb{1}(\boldsymbol{s}_t - \tau); \\ \hat{\boldsymbol{y}}_t^v = \boldsymbol{m}_t \odot \boldsymbol{y}^{v\cup a}, \end{cases} \tag{4}$$

其中，当输入 $x_i \geq 0$ 时，$\mathbb{1}(x_i)$输出“1”，否则输出“0”；$i = 1, 2, ..., C$，且 $\boldsymbol{m}_t \in \mathbb{R}^{1\times C}$。

这一伪标签生成过程可以轻松应用于所有视频片段。借助所提出的PLG，我们可以为每个视频获得片段级伪标签，记为$\hat{}\boldsymbol{Y}^v = \{\hat{}\boldsymbol{y}_t^v\} \in \mathbb{R}^{T\times C}$。注意，视频级伪标签$\hat{}\boldsymbol{y}^v \in \mathbb{R}^{1\times C}$ $\boldsymbol{Y}^v$ $y_c^v = \mathbb{1}$可以从$\hat{}$轻松获得，其中$\hat{}(\sum_{t=1}^{T} \hat{}\boldsymbol{Y}_{t,c}^v)$表示：如果某一事件类别存在于至少一个视频片段中，则该类别包含在视频级标签中。

B. 丰富度感知损失

在这一部分中，由于我们已经获得了片段级伪标签$\hat{}\boldsymbol{Y}^v$，因此尝试从片段级对模型进行正则化。这对于弱监督音视频事件定位任务至关重要，因为该任务需要预测每个片段的结果，但只提供视频级标签。为了更好地监督模型，我们从两个方面利用获得的伪标签。

基本视频级损失。现有方法通常采用式（1）中给出的目标函数进行模型训练 [42], [45], [48], [49], ，其中 $\boldsymbol{y}^v \in \mathbb{R}^{1\times C}$ 是通过标签平滑获得的视频级标签。相反，我们使用 PLG 模块生成的视频级伪标签 $\hat{}\boldsymbol{y}^v \in \mathbb{R}^{1\times C}$ 作为新的监督信号。于是，目标函数更新为：

$$\mathcal{L}_{\mathcal{V}} = \mathcal{L}_{\text{bce}}(\boldsymbol{p}^{v\cup a}, \boldsymbol{y}^{v\cup a}) + \mathcal{L}_{\text{bce}}(\boldsymbol{p}^a, \overline{\boldsymbol{y}}^a) + \mathcal{L}_{\text{bce}}(\boldsymbol{p}^v, \hat{\boldsymbol{y}}^v). \tag{5}$$

片段级丰富度感知损失。受以下观察启发，我们提出了一种新的丰富度感知损失。**1)** 逐片段伪标签的每一行，例如 例如，$\hat{}\boldsymbol{Y}_{t,:}^v \in \mathbb{R}^{1\times C}$，表示所有视觉事件是否都出现在第 t 个片段中。例如，如图2-2所示，第一个片段中有三个视觉事件，即，狗、公鸡、$\hat{}$和语音，$\boldsymbol{Y}_{1,:}^v = [1, 1, 1]$。而最后两个片段中仅$\hat{}\hat{}$包含一个狗 事件，即，$\boldsymbol{Y}^{v(T-1)}_{,:} = \boldsymbol{Y}_{T,:}^v = [1, 0, 0]$。这反映了不同片段中事件类别的丰富程度，它表示每个视频片段中存在多少种事件类别。我们将第 t 个片段 cr_t的类别丰富度定义为第 t 个片段的类别数与整个视频的类别数之比，记作：

$$cr_t = \frac{\sum_{c=1}^{C} \hat{\boldsymbol{Y}}_{t,c}^v}{\sum_{c=1}^{C} \boldsymbol{y}_c^{v\cup a}} \tag{6}$$

因此，我们可以得到所有片段的类别丰富度向量 $\boldsymbol{cr} \in \mathbb{R}^{T\times 1}$。在图2-2所示的示例中，第一个和最后一个片段的类别丰富度，即即，cr_1 和cr_T，分别为 1 和 1/3。

2) 类似地，伪标签$\hat{}\ \boldsymbol{Y}_{:,c}^v \in \mathbb{R}^{T\times 1}$ 的每一列表示包含以下事件的片段数量：c类。我们将第 c 类的片段丰富度 sr_c 定义为包含该类别的片段数与所有视频片段数之比，公式如下：

$$sr_c = \frac{1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{Y}}_{t,c}^v. \tag{7}$$

推广到所有 C 个事件类别后，我们可以得到所有类别的片段丰富度向量 $\boldsymbol{sr} \in \mathbb{R}^{1\times C}$。在图2-2所示的示例中，事件类别狗和语音，即，$sr_1$ 和 sr_3 的片段丰富度分别为 1 和 1/4。

$\hat{}$ 给定片段级伪标签 $\boldsymbol{Y}^v$，我们可以获得类别丰富度 $\boldsymbol{cr}$ 和片段丰富度 $\boldsymbol{sr}$。利用基线网络的视觉预测 $\boldsymbol{P}^v \in \mathbb{R}^{T\times C}$，我们可以用相同的方式计算类别丰富度和片段丰富度，分别记为 $\boldsymbol{pcr} \in \mathbb{R}^{T\times 1}$ 和 $\boldsymbol{psr} \in \mathbb{R}^{1\times C}$。随后，我们提出丰富度感知损失 $\mathcal{L}_{\mathcal{S}}$，以对齐视觉预测与伪标签的丰富度，其计算方式为：

$$\mathcal{L}_{\mathcal{S}} = \mathcal{L}_{\text{bce}}(\boldsymbol{pcr}, \boldsymbol{cr}) + \mathcal{L}_{\text{bce}}(\boldsymbol{psr}, \boldsymbol{sr}). \tag{8}$$

本文中 AVVP 的总目标函数 $\mathcal{L}_{\text{total}}$ 由基础损失 $\mathcal{L}_{\mathcal{V}}$ 和丰富度感知损失 $\mathcal{L}_{\mathcal{S}}$ 组成，即：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathcal{V}} + \lambda\mathcal{L}_{\mathcal{S}}, \tag{9}$$

其中，λ 是一个权重参数。

C. 伪标签去噪

PLG 能够生成可靠的伪标签，但仍存在受预训练技术限制的噪声。例如，当片段中的主体被部分遮挡或光照不足时，PLG 很难准确识别发生了哪些事件。伪标签的质量是现有标签去噪方法 [48], [49]中不可避免的常见问题。因此，我们设计了一种伪标签去噪（PLD）策略，以优化由 PLG 获得的伪标签。

具体而言，我们首先使用式9所示的目标函数训练一个基线模型。然后，针对训练数据，计算预测 $\boldsymbol{P}^v$ 与伪标签 $\boldsymbol{Y}^v$ 之间逐元素的前向传播损失矩阵 $\boldsymbol{\mathcal{M}}$between$\hat{}$: $\boldsymbol{\mathcal{M}} = \mathcal{L}_{\text{bce}}(\boldsymbol{P}^v, \hat{}\boldsymbol{Y}^v)$，以及$\boldsymbol{\mathcal{M}} \in \mathbb{R}^{T\times C}$。将$\boldsymbol{\mathcal{M}}$ 的第 j列记为$\boldsymbol{\mathcal{M}}_{\cdot j} \in \mathbb{R}^{T\times 1}$，它表示特定第 j个事件类别中所有片段的交叉熵值。预训练基线模型通常会在伪标签与真实标注相同的片段上产生较小的损失值。反之，伪标签错误的片段通常具有较大的损失，因此应反转其伪标签。

请注意，视频级伪标签$\hat{}\boldsymbol{y}^v \in \mathbb{R}^{1\times C}$ 表示视觉模态中存在的事件类别。我们使用$\hat{}\boldsymbol{y}^v$ 对矩阵$\boldsymbol{\mathcal{M}}$ 进行掩码，以确保依据预测的事件类别进行标签去噪。对于事件类别词汇表中视频里未出现的其他事件类别，将$\boldsymbol{\mathcal{M}}$中相应值设为零，以放宽其伪标签。对于图2-2中的示例，我们只需对前三列的伪标签进行去噪，这三列分别对应$\hat{}\boldsymbol{y}^v$中的狗、公鸡 和语音 事件类别。掩码矩阵$\boldsymbol{\mathcal{M}'}$ 的计算方式如下:

$$\boldsymbol{\mathcal{M}'} = f_{\text{rpt-T}}(\hat{\boldsymbol{y}}^v) \odot \boldsymbol{\mathcal{M}} \tag{10}$$

其中$\boldsymbol{\mathcal{M}'} \in \mathbb{R}^{T\times C}$，且 $f_{\text{rpt-T}}(\hat{}\boldsymbol{y}^v)$表示沿时间维度将$\hat{}\boldsymbol{y}^v$ 重复 T 次的操作，而$f_{\text{rpt-T}}(\hat{}\boldsymbol{y}^v) \in \mathbb{R}^{T\times C}$。

对于第 j类，所有视频片段的前向损失为$\boldsymbol{\mathcal{M}}'_{\cdot j} \in \mathbb{R}^{T\times 1}$。损失值异常大的片段可能带有不正确的伪标签。为了纠正此类片段的伪标签，我们将$\boldsymbol{\mathcal{M}}'_{\cdot j}$（可信预测分数）中损失值最小的前$K$ 个的平均值设为阈值 μ_j。通过将每个视频片段的交叉熵值与 μ_j进行比较，我们可以获得二值掩码向量 $\boldsymbol{\varphi}_j \in \mathbb{R}^{T\times 1}$，其中**“1”**表示该片段的损失大于 μ_j，并且可能带有不正确的伪标签。该过程可表示为:

$$\begin{cases} \mu_j = f_{\text{avg}}(f_{\mathbb{k}}(\boldsymbol{\mathcal{M}}'_{\cdot j})) \cdot \alpha, \\ \boldsymbol{\varphi}_j = \mathbb{1}(\boldsymbol{\mathcal{M}}'_{\cdot j} - \mu_j), \end{cases} \tag{11}$$

其中， $f_{\mathbb{k}}$ 和 f_{avg} 分别表示损失值最小的前K 个的选择操作和平均操作。 α 是用于放大平均损失的缩放因子。

扩展到所有事件类别后，我们可以得到$\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_j\} \in \mathbb{R}^{T\times C}$。最后，可以通过反转由 $\boldsymbol{\Phi}$ 反映出的具有大损失值的位置，对 PLG 生成的伪标签 $\hat{}\boldsymbol{Y}^v$ 进行细化，记为 $\widetilde{\boldsymbol{Y}}^v = f_{\sim}(\hat{}\boldsymbol{Y}^v, \boldsymbol{\Phi})$。如图2-2所示，对于事件类别 公鸡，伪 PLG 生成的标签在最后两个片段中为**“0”**，表示这些片段中没有出现公鸡 。然而，与其他片段相比，这两个片段的前向损失较大（图2-3中的紫色方框标出）。这表明事件类别公鸡 在这两个片段中的伪标签不正确，因此会在去噪过程中被反转。经 PLD 细化的伪标签可用作模型训练的新监督。

V. 实验

A. 实验设置

数据集。 实验在公开的 *Look, Listen, and Parse*（*LLP*） [40] 数据集上进行。该数据集包含 11,849 个视频，涵盖 25 个常见的视听类别，例如人类、动物、车辆和乐器等场景。每个视频时长为 10 秒，约 61% 的视频包含多个事件类别。LLP 数据集中的视频划分为 10,000 个训练视频、649 个验证视频和 1,200 个测试视频。训练集仅提供视频级标签，即音频事件和视觉事件的标签并集。对于验证集和测试集，还额外提供了每个音频和视觉模态的分段事件标签。

评估指标。 遵循现有研究 [40], [45], [48], [49], ，我们通过衡量所有类型事件的解析结果来评估各方法，即音频事件（**A**）、视觉事件（**V**）和视听事件（**AV**，同时可听见且可看见）。我们采用片段级和事件级 F 分数作为评估指标。片段级指标通过将预测的事件与每个视频片段的真实标注进行比较，衡量预测事件的质量。事件级指标则将包含相同事件类别的连续片段视为一个完整事件，并以平均交并比 = 0.5 作为阈值计算 F 分数。三类事件的平均解析结果记为**“Type@AV”**指标。此外，还有一个**“Event@AV”**指标，同时考虑音频事件和视觉事件的预测结果。

实现细节。*1*）特征提取。对于 LLP 数据集，每个视频被划分为 10 个连续的 1 秒片段。为进行公平比较，我们采用相同的特征提取器来提取音频和视觉特征。具体而言，使用在 AudioSet 数据集 [76]上预训练的 VGGish [1] 网络提取 128 维音频特征。使用预训练的 ResNet 152 [77] 和 R(2+1)D [78] 分别提取 2D 和 3D 视觉特征。低层视觉特征由 2D 和 3D 视觉特征拼接而成。在伪标签生成过程中使用基于ViT的CLIP [75]，并冻结其参数。*2*）训练过程。对于 LLP 数据集训练集中的每个视频，我们首先使用 PLG 模块离线生成逐片段视觉伪标签。然后，使用式 9 所示的目标函数 $\mathcal{L}_{\text{total}}$ 训练基线模型 HAN [40]。式 9 中用于平衡视频级和片段级损失的超参数λ 通过经验设为 0.5。随后，将该预训练模型用于我们的 PLD，以进一步优化伪标签。最后，使用 PLD 得到的伪标签再次监督基线模型训练。对于所有训练过程，

表 I**参数研究：τ 以及PLG 中使用的提示词**。采用不同设置生成片段级视觉伪标签，因此也可用于获取视频级视觉伪标签。本表报告视频级视觉轨道中伪标签与真实标注之间的精确率、召回率和平均值（%）。**‘-’表示直接将视频标签指定为视觉事件标签所得到的结果。**片段级性能将在表 II 中讨论。本实验在 LLP 数据集的验证集上进行。

参数设置		精确率（V）	召回率（V）	平均值
τ	提示词			
-	-	66.96	87.88	77.42
0.035	P1	83.69	81.15	82.42
0.040		83.77	81.15	**82.46**
0.041		81.33	77.38	79.35
0.043		38.29	34.14	36.22
0.040	**P1**	83.77	81.15	**82.46**
	P2	80.12	75.46	77.79
	P3	78.22	72.74	75.48
	P4	83.08	79.66	81.37

表 II**在中使用的 K 和 α 的参数研究PLD**。针对逐片段伪标签去噪测试了 K 和 α 的不同取值。报告了去噪后的视觉伪标签的片段级和事件级 F 分数。值得注意的是，在本表中，我们评估了视觉轨道（V）的伪标签。最后一列为平均值。**“-”表示未经标签去噪、由 *PLG* 生成的伪标签结果。**本实验在 LLP 数据集的验证集上进行。

参数设置		片段。（V）	事件。（V）	平均值
K	α			
-	-	70.29	64.68	67.49
2	10	72.45	67.13	69.79
3		72.26	67.30	69.78
5		**72.51**	**68.09**	**70.30**
6		71.84	66.91	69.38
5	5	72.09	66.44	69.26
	10	**72.51**	**68.09**	**70.30**
	20	72.05	67.58	69.81

我们采用 Adam 优化器训练模型，迷你批次大小为 32，学习率为 3×10^{-4}。总训练轮数设为 30。所有实验均使用 PyTorch [79] ，在一块 NVIDIA GeForce-RTX-2080-Ti GPU 上进行。代码将会发布。

B. 参数研究

我们首先对方法中使用的关键参数进行研究，即 PLG 模块中的掩码阈值 τ 和 CLIP 提示词，以及 PLD 策略中的 top-K 和缩放因子 α 。本节实验在 LLP 数据集的验证集上进行，该验证集的片段级事件标签已知，可用于参数调优。因此，我们也会在本节讨论伪标签正确性的度量方式。

对 PLG 中 τ 和提示词的研究。τ 用作阈值，在掩码计算（公式 4）中筛选事件类别与视频片段之间余弦相似度的高分。为探究其影响，我们将 CLIP 提示词固定为 **P1** –“这张照片中包含 [CLS]”，并测试 τ 的多个取值，以生成片段级视觉伪标签，从而也能在视频级获取视觉伪标签。随后，我们在验证集上以视频级为单位计算伪标签与真实标注之间的精确率和召回率，并使用平均指标选出最佳设置。如表 I 上半部分所示，当 $\tau = 0$时，平均结果最佳。 $\tau \in [0.035, 0.04]$时，性能相对稳定；而当 τ 从 0.041 变为 0.043 时，精确率显著下降。我们认为，这种敏感性与公式 3 中的 *softmax* 操作有关，该操作会将相似度分数压缩为较小的 logits。

此外，我们还探讨了提示词在 CLIP 中的影响。具体而言，我们测试了四种类型的提示词，即，我们的默认**P1** ——“这张照片包含 [CLS]”，**P2**——“这张照片包含 [CLS] 的场景”，**P3** – “这张照片包含 [CLS]所表示的视觉场景”和**P4** – “这是一张[CLS]的照片”。我们使用这些不同的提示词生成伪标签，并将其与真实标注进行比较。如表 I 下半部分所示，使用这些不同提示词生成的伪标签基本保持一致。使用**P1**提示词时，伪标签的平均分数最高。因此，在后续实验中，我们将 τ 设为 0.04，并采用**P1** 提示词作为最佳设置。值得注意的是，在最佳设置下，视频级伪标签的精确率约为 84%。而直接将视频标签指定为视觉事件标签时，精确率仅为 ∼67%。这表明 PLG 能够令人满意地从弱视频标签中分离出视觉事件。伪标签的片段级评估将在表 II 中讨论。

研究 K 和 α 在 PLD 中的伪标签去噪过程中，对于每个预测的事件类别，沿时间维度选取前 K 个最小前向损失，并将其放大 α 倍后作为阈值，以确定哪些片段的伪标签需要修正。为探究这两个参数的影响，我们测试了它们的不同组合。我们使用视觉事件的片段级和事件级 *F* 分数来衡量去噪后伪标签的质量。表 II 中的结果表明，去噪后的伪标签比 PLG 生成的原始标签带来了更好的性能。最优设置为 $K = 5$和 $\alpha = 10$。在最优设置下，验证集的视觉伪标签在片段级和事件级上的 F 分数分别为 72.51% 和 68.09%。由于训练集未提供片段级标签，我们无法准确计算其 F 分数。不过，我们推测训练集的 F 分数应该不低，甚至可能优于验证集（片段级和事件级 F 分数分别为 72.51% 和 68.09%），因为这两个集合来自同一个数据集，具有相似的数据分布，而且我们使用相同的 PLG 模块获取伪标签。值得注意的是，在测试集上，目前最先进的视觉事件解析性能在片段级达到 58.3% 的 F 分数，

表 III**消融主要模块的研究。** $\mathcal{L}_{\mathcal{S}}$ 表示所提出的丰富度感知损失（公式 8），$\mathcal{L}'_{\mathcal{S}}$ 表示一种原生损失，仅计算预测与伪标签之间的二元交叉熵损失。我们报告了 LLP 数据集测试集上的结果。

Id	主要模块			片段级					事件级				
	PLG	RL	PLD	A	V	AV	Type@AV	Event@AV	A	V	AV	类型@视听	事件@视听
①	✘	✘	✘	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
②	✔	✘	✘	60.0	63.7	57.3	60.3	58.8	51.1	60.3	50.4	53.9	50.2
③	✔	✔-$\mathcal{L}'_{\mathcal{S}}$	✘	60.2	64.1	57.6	60.6	59.0	50.8	60.5	50.1	53.8	49.8
④	✔	✔-$\mathcal{L}_{\mathcal{S}}$	✘	60.5	64.8	58.3	61.2	59.4	51.4	61.5	51.2	54.7	50.8
⑤	✔	✔	✔	**61.1**	**65.5**	**59.3**	**62.0**	**60.1**	**52.1**	**62.4**	**52.4**	**55.6**	**51.3**

表IV**COMPARISON WITH THE STATE-OF-THE-ARTS.** ♢ 所有这些方法均致力于为视听视频解析任务生成更好的伪标签，并且都是基于基线 HAN [40] 框架开发的。结果在 LLP 数据集的测试集上报告。

方法	片段级					事件级				
	A	V	AV	类型@视听	事件@视听	A	V	AV	类型@视听	事件@视听
视听事件 [25]	47.2	37.1	35.4	39.9	41.6	40.4	34.7	31.6	35.5	36.5
视听时空动态网络 [26]	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6	37.7
HAN [40]	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
MM-Pyramid [45]	60.9	54.4	50.0	55.1	57.6	52.7	51.8	44.4	49.9	50.5
MGN [42]	60.8	55.4	50.4	55.5	57.2	51.1	52.4	44.4	49.3	49.1
交叉视图跨模态检索系统 [46]	59.2	59.9	53.4	57.5	58.1	51.3	55.5	46.2	51.0	49.7
DHHN [47]	61.3	58.3	52.9	57.5	58.1	54.0	55.1	47.3	51.5	51.5
♢MA [48]	60.3	60.0	55.1	58.9	57.9	53.6	56.4	49.0	53.0	50.6
♢JoMoLD [49]	**61.3**	63.8	57.2	60.8	59.9	**53.9**	59.9	49.6	54.5	**52.5**
♢**VPLAN（我们的方法）**	61.1	**65.5**	**59.3**	**62.0**	**60.1**	52.1	**62.4**	**52.4**	**55.6**	51.3

由方法 DHHN [47] 实现。这一差距表明，PLG 生成并经 PLD 优化的训练集伪标签可作为模型学习的可靠目标。

C. 消融研究

在本节中，我们通过一些消融研究来探究我们方法中各模块的影响。实验结果如表 III 所示。id-① 所在的行表示基线 HAN [40]的性能。

PLG 的影响。借助 PLG 模块，我们获得了片段级和视频级伪标签，分别记为 $\hat{\boldsymbol{Y}}^v$ 和 $\hat{\boldsymbol{y}}^v$。值得注意的是，id-①（基础 HAN）仅使用视频级伪标签 $\boldsymbol{y}^v$ 进行模型训练（式 2）。为公平比较，我们使用视频级伪标签 $\hat{\boldsymbol{y}}^v$ 监督模型训练（式 5）。如表 III 中 row-② 所示，利用 PLG 生成的伪标签显著提升了视觉相关事件预测的性能。片段级的平均指标 Type@AV 从 54.0% 提升至 60.3%，而事件级则从 47.7% 提升至 53.9%。这些结果表明，所提出的 PLG 设计有效，能够从大型预训练模型中成功借鉴免费的可靠先验知识（伪标签）。

丰富度感知损失（RL）的影响。公式 8 中的损失 $\mathcal{L}_{\mathcal{S}}$ 旨在利用片段级伪标签，作为视频级监督的补充。为了验证其有效性，我们将其与一种直接计算预测和伪标签之间二元交叉熵损失的原生变体进行比较，记为 $\mathcal{L}'_{\mathcal{S}} = \mathcal{L}_{\text{bce}}(\boldsymbol{P}^v, \hat{\boldsymbol{Y}}^v)$。如表 III 的第 ③ 行和第 ④ 行所示，所提出的 RL $\mathcal{L}_{\mathcal{S}}$比 $\mathcal{L}'_{\mathcal{S}}$ 更有助于视频事件解析。例如，片段级和事件级的 Type@AV 指标均提升了约 1%。原因可能是，$\mathcal{L}'_{\mathcal{S}}$ 强制对齐每个片段的预测和伪标签。相比之下，$\mathcal{L}_{\mathcal{S}}$ 提供软监督，通过感知丰富度信息来优化模型。

PLD 的影响。 PLD 的影响可以从两个方面观察。一方面，PLD 能够提供比 PLG 更准确的伪标签。如表 II 所示，在验证集上，PLG 的平均 F分数为 67.49%，而 PLD 为 70.30%。另一方面，PLD 比 PLG 更有助于模型训练。我们以 PLD 替代 PLG 作为新的监督来训练 HAN 模型。如表 III 第-⑤ 行所示，使用 PLD 时，模型在大多数指标上都取得了更优的表现。这表明 PLD 获得的伪标签比 PLG 的更准确，因此能够提供好得多的监督。此外，这也验证了 PLD 中标签去噪策略的有效性。

D. 与最先进方法的比较

我们采用第 V-B 节中研究得出的最优参数运行 VPLAN 方法，并报告其在 LLP 数据集测试集上的表现。与现有方法的比较结果

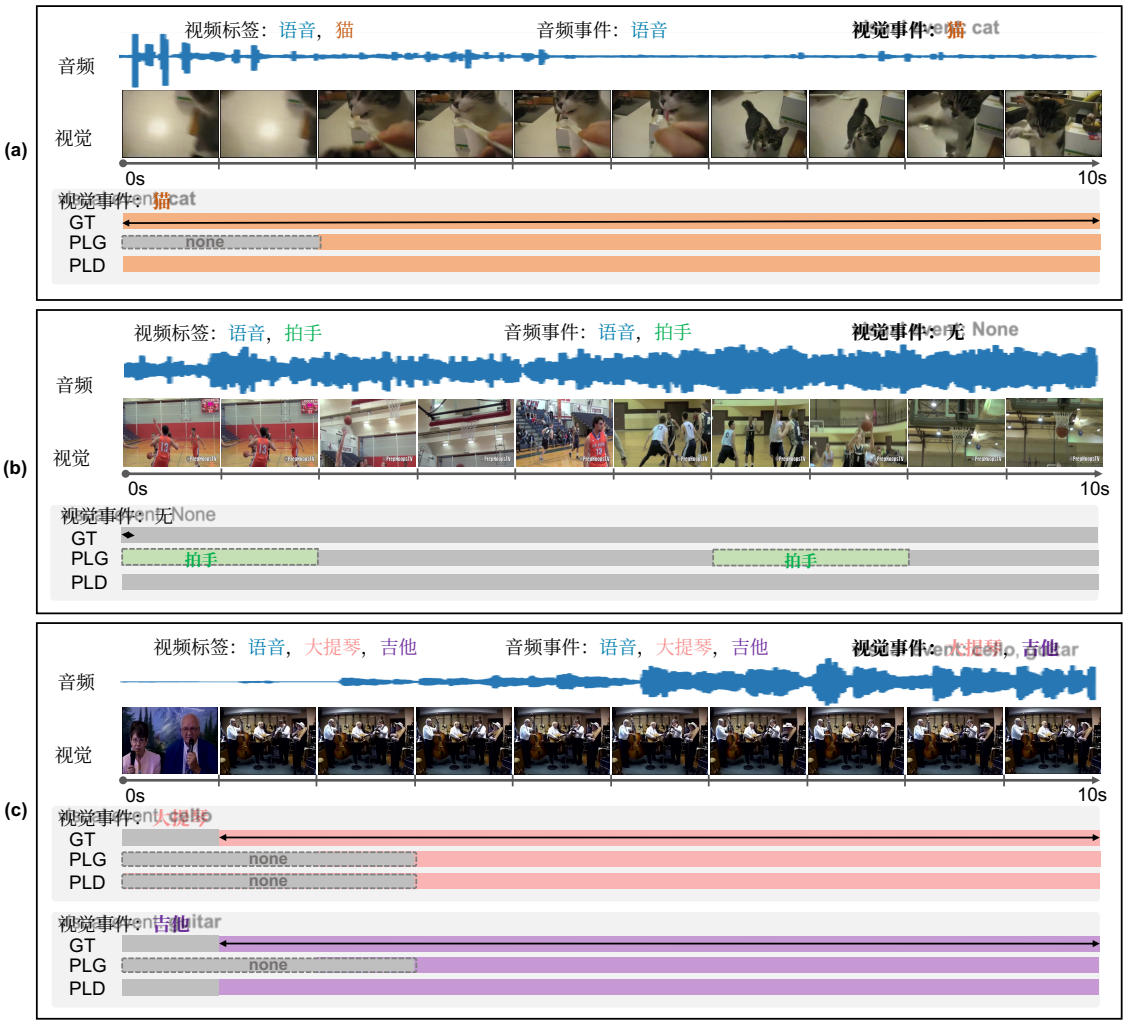


图3. **伪标签的可视化**。“真实标注”表示真实标注。“PLG”和“PLD”分别表示通过 PLG 和 PLD 获得的伪标签。PLG 基本上将视觉事件与弱视频标签解耦，并提供分段事件类别，而 L-PLD 则进一步修正伪标签。

表 V**G我们方法的泛化能力**。我们的方法能够生成可靠的伪标签，可用于视听视频解析任务中的其他方法。我们评估了两个最先进的网络，即 MM-PYRAMID [45] 和 MGN [42]。由我们的 PLG 生成并经我们的 PLD 优化的伪标签持续提升了这些模型的性能，也优于现有方法 MA [48]。* 表示这些结果是我们使用官方源代码复现的。

方法	片段级					事件级				
	A	V	AV	类型@视听事件@视听		A	V	AV	类型@视听事件@视听	
*MM-Pyramid [45]	60.5	52.3	48.1	53.6	57.5	51.5	48.3	42.5	47.4	48.5
*MM-Pyramid + 多模态对齐 [48]	58.6	**57.3**	52.1	56.0	57.2	50.5	**53.1**	45.9	49.9	48.5
MM-Pyramid **+PLG**	59.8	**63.9**	57.3	60.3	59.1	50.0	**60.3**	50.2	53.5	48.6
MM-Pyramid **+PLD**	60.1	**64.8**	58.0	61.0	59.5	51.4	**60.7**	50.7	54.2	49.7
MGN [42]	60.8	55.4	50.4	55.5	57.2	51.1	52.4	44.4	49.3	49.1
MGN + 多模态对齐 [48]	60.2	**61.9**	55.5	59.2	58.7	50.9	**59.7**	49.6	53.4	49.9
MGN **+PLG**	60.3	**63.0**	56.4	59.9	59.1	50.5	**61.1**	50.4	54.0	49.5
MGN **+PLD**	59.6	**64.0**	56.9	60.2	58.9	49.8	**61.8**	50.5	54.0	49.3

见表IV。我们的方法整体表现更优。**首先**，与基线方法 HAN [40]，相比，我们的方法显著提升了性能，尤其是在视觉相关指标方面。以视觉事件预测为例（V

以表中）为例，片段级指标从 52.9% 提升至 65.5%（↑ 12.6%），事件级指标从 48.9% 提升至 62.4%（↑ 13.5%）。**其次**，我们的方法在生成

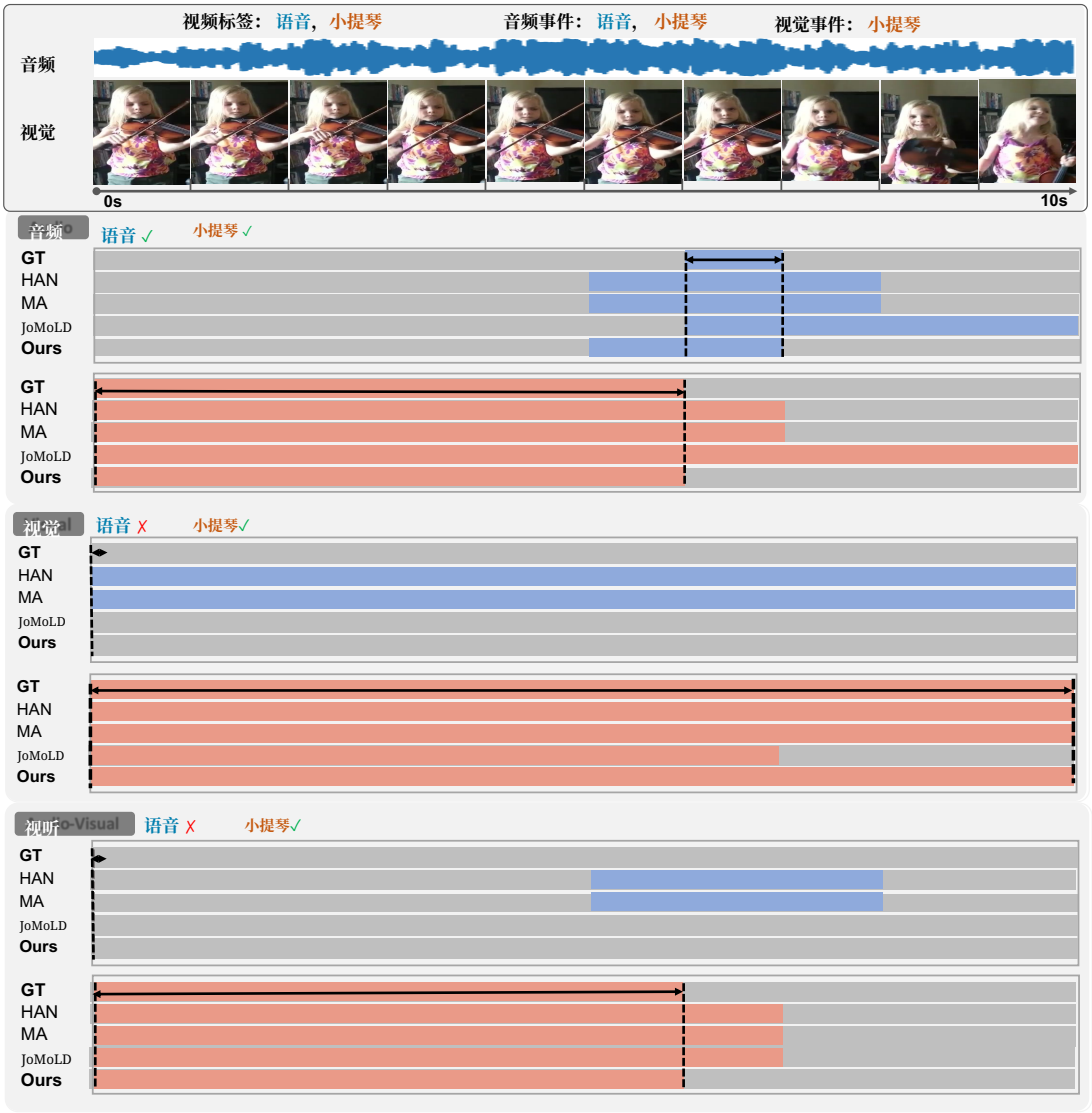


图4. **视听视频解析的定性示例**。 我们将我们的方法与 HAN [40], MA [48] 和 JoMoLD [49]进行比较。“GT”表示真实标注。我们的方法成功识别出视觉轨道中只有一个事件小提琴，并且在解析音频事件和视听事件方面也更加准确。

用于 AVVP 任务的伪标签方面也优于其他方法。如表格下半部分所示，我们的方法在音频事件解析方面表现相当，但在视觉事件解析和视听事件解析方面，整体上分别超过方法 MA [48] 个百分点和方法 JoMoLD [49] 个百分点。这些结果证明了我们整体方法设计的优越性。

E. 我们方法的泛化能力

我们方法的一项核心贡献是能够提供高质量的伪标签，从而更好地指导模型优化。该方法还可以直接应用于现有的视听视频解析任务方法。为探究其影响，我们考察了最近提出的两种方法，即., MM-Pyramid [45]和 MGN [42]。具体而言，我们分别使用由我们的 PLG 生成、并由我们的 PLD 优化后的伪标签来训练模型。实验结果如表 V 所示。PLG 和 PLD 都显著提升了基础模型，尤其是在视觉事件解析和视听事件解析方面。以 MM-Pyramid [45] 方法为例，使用 PLG 和 PLD 后，片段级视觉事件解析性能分别从 52.3% 提升至 63.9% 和 64.8%。PLD 的优势得益于所提出的标签去噪策略。MGN [42]也呈现出类似的提升。值得注意的是，与使用 MA [48]生成的伪标签相比，这两种模型使用我们的伪标签时表现也更好。这些结果进一步表明了我们方法的优越性和泛化能力。

F. 定性结果

伪标签和标签去噪的可视化。我们首先展示 PLG 和 PLD 生成的三个伪标签可视化示例。如图3 (a)所示，该视频包含语音 和猫这两个事件，而语音是视觉模态的干扰因素。PLG 成功

识别出视觉轨道中仅发生了猫 事件，并生成逐片段伪标签。PLG 生成的大多数伪标签都是正确的。然而，由于前两个片段中的物体过于模糊，猫 事件被错误识别。得益于提出的标签去噪策略，PLD 对其进行了修正。图3 (b)中没有视觉事件。由于播放器在这些片段中的动作较为复杂，PLG 错误地将少数片段分类为拍手 事件。PLD 也修正了这些错误预测。图3 (c)中，视觉轨道包含两个事件，即：大提琴 和吉他。PLG 和 PLD 都未能识别第二和第三个片段中的大提琴 事件，但 PLD 成功修正了这些片段中的吉他 预测。总体而言，PLG 生成的伪标签依赖于预训练 CLIP 提供的先验知识，而 PLD 则受益于额外的修正过程（——联合探索基线 HAN [40] 训练和前向损失 $\mathcal{L}_{\text{bce}}$计算，可以修正 PLG 中可能存在的错误伪标签）。

视听视频解析的可视化示例。我们还在图4中展示了一个定性的视频解析示例。我们将我们的方法与 HAN [40]，MA [48]，和 JoMoLD [49]进行了比较。MA 和 JoMoLD 均基于 HAN 开发，并尝试生成视频级伪标签，以改进模型训练。如图所示，视频中存在两个事件，即语音和小提琴，而视觉事件仅包含小提琴。在音频事件解析方面，所有方法都识别出音频轨道中发生了这两个事件，但我们的方法能更准确地定位时间片段。在视觉事件解析方面，HAN 和 MA 错误地预测语音和小提琴这两个事件都发生在视觉轨道中。尽管 JoMoLD 成功识别出视觉事件只有小提琴，但它错误地预测了最后三个视频片段。相比之下，我们的方法能够准确解析视觉事件。我们的方法在最终的视听事件解析中也表现更佳。这一优势源于我们的方法能够生成高质量的片段级伪标签，为模型训练提供更好的监督。

VI. 结论

我们提出了一种用于弱监督音视频解析任务的方法——视觉伪标签探索（VPLAN）。VPLAN 是该领域首次尝试生成片段级伪标签的方法，它首先通过伪标签生成模块，利用强大的 CLIP 在视觉模态中确定（片段级）事件，并将其作为伪标签。在这项工作中，我们研究了伪标签中包含的类别丰富度和片段丰富度，并提出了一种新的丰富度感知损失，用于约束标签对齐。最后，我们提出了一种伪标签去噪策略，以优化伪标签并更好地引导预测。在 LLP 数据集上的定性和定量实验结果证实，我们的方法能够生成并有效利用高质量伪标签。我们提出的所有技术均可直接供学界使用。

致谢

我们谨感谢 Liang Zheng 博士和 Yunfeng Diao 博士提出的建设性意见与建议。该工作获得了中国国家自然科学基金（72188101、61725203、62020106007 和 62272144）以及安徽省重大项目（202203a05020011）的资助。

参考文献

[1] S. Hershey、S. Chaudhuri、D. P. Ellis、J. F. Gemmeke、A. Jansen、R. C. Moore、M. Plakal、D. Platt、R. A. Saurous、B. Seybold 等，“用于大规模音频分类的CNN架构”，载于*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，2017年，第131–135页。[2] Q. Kong、Y. Xu、W. Wang和M. D. Plumbley，“采用注意力模型的Audio Set分类：一种概率视角”，载于*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，2018年，第316–320页。[3] A. Kumar、M. Khadkevich和C. Fügen，“利用卷积神经网络从弱标注音频中迁移知识，以识别声音事件和场景”，载于*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，2018年，第326–330页。[4] Y. Gong、Y.-A. Chung和J. Glass，“AST：音频频谱变换器”，*arXiv*预印本 *arXiv:2104.01778*，2021年。[5] A. Karpathy、G. Toderici、S. Shetty、T. Leung、R. Sukthankar和L. Fei- Fei，“使用卷积神经网络进行大规模视频分类”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集 *(CVPR)*，2014年，第1725–1732页。[6] X. Long、C. Gan、G. De Melo、J. Wu、X. Liu和S. Wen，“注意力簇：一种纯注意力式的视频分类局部特征整合方法”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集 *(CVPR)*，2018年，第7834–7843页。[7] X. Long、C. Gan、G. Melo、X. Liu、Y. Li、F. Li和S. Wen，“用于视频分类的多模态无键注意力融合”，载于*AAAI*人工智能会议论文集 *(AAAI)*，2018年，第1–8页。[8] D. Tran、H. Wang、L. Torresani和M. Feiszli，“使用通道分离卷积网络进行视频分类”，载于*IEEE*国际计算机视觉会议论文集 *(ICCV)*，2019年，第5552–5561页。[9] R. Zeng、W. Huang、M. Tan、Y. Rong、P. Zhao、J. Huang和C. Gan，“用于时序动作定位的图卷积网络”，载于*IEEE*国际计算机视觉会议论文集 *(ICCV)*，2019年，第7094–7103页。[10] Y.-W. Chao、S. Vijayanarasimhan、B. Seybold、D. A. Ross、J. Deng和R. Sukthankar，“重新思考用于时序动作定位的Faster R-CNN架构”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集 *(CVPR)*，2018年，第1130–1139页。[11] Z. Zhu、W. Tang、L. Wang、N. Zheng和G. Hua，“丰富时序动作定位的局部与全局上下文”，载于*IEEE*国际计算机视觉会议论文集 *(ICCV)*，2021年，第13 516–13 525页。[12] J. Gao、M. Chen和C. Xu，“用于弱监督时序动作定位的细粒度时序对比学习”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集 *(CVPR)*，2022年，第19 999–20 009页。[13] Y. Wei、D. Hu、Y. Tian和X. Li，“视听上下文中的学习：综述、分析与新视角”，*arXiv*预印本 *arXiv:2208 09579*，2022年。。[14] R. Arandjelovic和A. Zisserman，“观看、聆听与学习”，载于*IEEE*国际计算机视觉会议论文集 *(ICCV)*，2017年，第609–617页。[15] ——，“会发声的物体”，载于欧洲计算机视觉会议论文集 *(ECCV)*，2018年，第435–451页。[16] A. Rouditchenko、H. Zhao、C. Gan、J. McDermott和A. Torralba，“自监督视听协同分割”，载于*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2019年，第2357–2361页。[17] A. Senocak、T.-H. Oh、J. Kim、M.-H. Yang和I. S. Kweon，“学习定位视觉场景中的声源”，载于*IEEE*计算机视觉与模式识别会议论文集 *(CVPR)*，2018年，第4358–4366页。[18] D. Hu、R. Qian、M. Jiang、X. Tan、S. Wen、E. Ding、W. Lin和D. Dou，“通过自监督视听匹配定位具有判别性的发声物体”，神经信息处理系统进展 *(NeurIPS)*，第10 077–10 087页，2020年。[19] D. Hu、F. Nie和X. Li，“用于无监督视听学习的深度多模态聚类”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集 *(CVPR)*，2019年，第9248–9257页。

[20] R. Qian、D. Hu、H. Dinkel、M. Wu、N. Xu和W. Lin，“从粗到细的多个声源定位”，载于《欧洲计算机视觉会议论文集 (ECCV)》，2020年，第292–308页。[21] H. Zhao、C. Gan、A. Rouditchenko、C. Vondrick、J. McDermott和A. Torralba，“像素之声”，载于《欧洲计算机视觉会议论文集 (ECCV)》，2018年，第570–586页。[22] T. Afouras、A. Owens、J. S. Chung和A. Zisserman，“从视频中自监督学习视听对象”，载于《欧洲计算机视觉会议论文集 (ECCV)》，2020年，第208–224页。[23] J. Zhou、J. Wang、J. Zhang、W. Sun、J. Zhang、S. Birchfield、D. Guo、L. Kong、M. Wang和Y. Zhong，“音视分割”，载于《欧洲计算机视觉会议论文集 (ECCV)》，2022年，第386–403页。[24] J. Zhou、X. Shen、J. Wang、J. Zhang、W. Sun、J. Zhang、S. Birchfield、D. Guo、L. Kong、M. Wang等，“结合语义的音视分割”，arXiv预印本 arXiv:2301.13190，2023年。[25] Y. Tian、J. Shi、B. Li、Z. Duan和C. Xu，“非受限视频中的视听事件定位”，载于《欧洲计算机视觉会议论文集 (ECCV)》，2018年，第247–263页。[26] Y.-B. Lin、Y.-J. Li和Y.-C. F. Wang，“用于视听事件定位的双模态序列到序列网络”，载于《IEEE国际声学、语音与信号处理会议 (ICASSP)》，2019年，第2002–2006页。[27] Y. Wu、L. Zhu、Y. Yan和Y. Yang，“用于视听事件定位的双重注意力匹配”，载于《IEEE国际计算机视觉会议论文集 (ICCV)》，2019年，第6292–6300页。[28] H. Xu、R. Zeng、Q. Wu、M. Tan和C. Gan，“用于视听事件定位的跨模态关系感知网络”，载于《ACM 国际多媒体会议论文集 (ACM MM)》，2020年，第3893–3901页。[29] H. Xuan、Z. Zhang、S. Chen、J. Yang和Y. Yan，“用于时间不一致音视事件定位的跨模态注意力网络”，载于《AAAI人工智能会议论文集 (AAAI)》，2020年，第279–286页。[30] J. Zhou、L. Zheng、Y. Zhong、S. Hao和M. Wang，“沿音视事件线进行正样本传播”，载于《IEEE/CVF计算机视觉与模式识别会议论文集 (CVPR)》，2021年，第8436–8444页。[31] T. Mahmud和D. Marculescu，“Ave-clip：基于Audioclip的视听事件定位多窗口时序变换器”，2022年，第1–10页。[32] V. Rao、M. I. Khalil、H. Li、P. Dai和J. Lu，“用于视听事件定位的双视角网络”，载于《欧洲计算机视觉会议论文集 (ECCV)》，2022年，第689–704页。[33] Y. Xia和Z. Zhao，“用于视听事件定位的跨模态背景抑制”，载于《IEEE/CVF计算机视觉与模式识别会议论文集 (CVPR)》，2022年，第19 989–19 998页。[34] Y. Wu、X. Zhang、Y. Wang和Q. Huang，“基于跨度的音视定位”，载于《ACM 国际多媒体会议论文集 (ACM MM)》，2022年，第1252–1260页。[35] J. Zhou、D. Guo和M. Wang，“沿音视事件线进行对比式正样本传播”，《IEEE模式分析与机器智能汇刊》，2023年。[36] H. Wang、Z.-J. Zha、L. Li、X. Chen和J. Luo，“用于视听事件定位的语义与关系调制”，《IEEE模式分析与机器智能汇刊》，2022年。[37] H. Yun、Y. Yu、W. Yang、K. Lee和G. Kim，“Pano-avqa：360度视频中的基于视觉依据的音视问答”，载于《IEEE/CVF国际计算机视觉会议论文集 (ICCV)》，2021年，第2031–2041页。[38] G. Li、Y. Wei、Y. Tian、C. Xu、J.-R. Wen和D. Hu，“学习在动态音视场景中回答问题”，载于《IEEE/CVF计算机视觉与模式识别会议论文集 (CVPR)》，2022年，第19 108–19 118页。[39] P. Yang、X. Wang、X. Duan、H. Chen、R. Hou、C. Jin和W. Zhu，“Avqa：用于视频视听问答的数据集”，载于《第30届ACM国际多媒体会议 (ACM MM)论文集》，2022年，第3480–3491页。[40] Y. Tian、D. Li和C. Xu，“统一多感官感知：弱监督视听视频解析”，载于《欧洲计算机视觉会议论文集 (ECCV)》，2020年，第436–454页。[41] C. Szegedy、V. Vanhoucke、S. Ioffe、J. Shlens和Z. Wojna，“重新思考用于计算机视觉的Inception架构”，载于《IEEE/CVF计算机视觉与模式识别会议论文集 (CVPR)》，2016年，第2818–2826页。[42] S. Mo和Y. Tian，“用于弱监督视听视频解析的多模态分组网络”，载于《神经信息处理系统进展 (NeurIPS)》，2022年。

[43] P. S. Pasi、S. Nemani、P. Jyothi 和 G. Ramakrishnan，《探究视听视频解析中的模态偏差》，arXiv预印本 arXiv:2203.16860，2022年。[44] J. Lamba、J. Akula、R. Dabral、P. Jyothi、G. Ramakrishnan 等，《用于视听视频解析的跨模态学习》，arXiv预印本 arXiv:2104.04598，2021年。[45] J. Yu、Y. Cheng、R.-W. Zhao、R. Feng 和 Y. Zhang，《MM-Pyramid：用于视听事件定位与视频解析的多模态金字塔注意力网络》，载于 ACM 国际多媒体会议（ACM MM）论文集，2022年，第6241–6249页。[46] Y.-B. Lin、H.-Y. Tseng、H.-Y. Lee、Y.-Y. Lin 和 M.-H. Yang，《探索用于弱监督视听视频解析的跨视频与跨模态信号》，《神经信息处理系统进展》（NeurIPS），第11 449–11 461页，2021年。[47] X. Jiang、X. Xu、Z. Chen、J. Zhang、J. Song、F. Shen、H. Lu 和 H. T. Shen，《DHHN：用于弱监督视听视频解析的双层级混合网络》，载于 ACM 国际多媒体会议（ACM MM）论文集，2022年，第719–727页。[48] Y. Wu 和 Y. Yang，《探索用于弱监督视听视频解析的异构线索》，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2021年，第1326–1335页。[49] H. Cheng、Z. Liu、H. Zhou、C. Qian、W. Wu 和 L. Wang，《用于弱监督视听视频解析的联合模态标签去噪》，载于欧洲计算机视觉会议（ECCV）论文集，2022年。[50] J. Deng、W. Dong、R. Socher、L.-J. Li、K. Li 和 L. Fei-Fei，《ImageNet：大规模层级图像数据库》，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2009年，第248–255页。[51] J.-B. Alayrac、J. Donahue、P. Luc、A. Miech、I. Barr、Y. Hasson、K. Lenc、A. Mensch、K. Millican、M. Reynolds 等，《Flamingo：用于少样本学习的视觉语言模型》，arXiv预印本 arXiv:2204.14198，2022年。[52] C. Jia、Y. Yang、Y. Xia、Y.-T. Chen、Z. Parekh、H. Pham、Q. Le、Y.-H. Sung、Z. Li 和 T. Duerig，《利用噪声文本监督扩展视觉与视觉语言表征学习》，载于国际机器学习会议（ICML），2021年，第4904–4916页。[53] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等，《通过自然语言监督学习可迁移视觉模型》，载于国际机器学习会议（ICML），2021年，第8748–8763页。[54] D. Arpit、S. Jastrzebski、N. Ballas、D. Krueger、E. Bengio、M. S. Kanwal、T. Maharaj、A. Fischer、A. Courville、Y. Bengio 等，《深入探究深度网络中的记忆化》，载于国际机器学习会议，2017年，第233–242页。[55] P. Hu、X. Peng、H. Zhu、L. Zhen 和 J. Lin，《利用噪声标签学习跨模态检索》，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2021年，第5403–5413页。[56] Y. Kim、J. M. Kim、Z. Akata 和 J. Lee，《大损失对弱监督多标签分类至关重要》，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2022年，第14 156–14 165页。[57] I. Z. Yalniz、H. Jégou、K. Chen、M. Paluri 和 D. Mahajan，《面向图像分类的十亿级半监督学习》，载于 arXiv预印本 arXiv:1905.00546，2019年。[58] Q. Xie、M.-T. Luong、E. Hovy 和 Q. V. Le，《使用噪声学生进行自训练，提升ImageNet分类性能》，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2020年，第10 687–10 698页。[59] H. Pham、Z. Dai、Q. Xie 和 Q. V. Le，《元伪标签》，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2021年，第11 557–11 568页。[60] M. N. Rizve、K. Duarte、Y. S. Rawat 和 M. Shah，《为伪标签方法辩护：面向半监督学习的不确定性感知伪标签选择框架》，arXiv预印本 arXiv:2101 20，2021年06329，1. 第–页。[61] B. Zoph、G. Ghiasi、T.-Y. Lin、Y. Cui、H. Liu、E. D. Cubuk 和 Q. Le，《重新思考预训练与自训练》，《神经信息处理系统进展》（NeurIPS），第3833–3845页，2020年。[62] Z. Hu、Z. Yang、X. Hu 和 R. Nevatia，《Simple：用于半监督分类的相似伪标签利用》，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2021年，第15 099–15 108页。[63] J. Kahn、A. Lee 和 A. Hannun，《端到端语音识别的自训练》，载于 ICASSP 2020-2020 IEEE声学国际会议

’语音与信号处理（*ICASSP*） ，2020，第 7084–7088 页。

[64] D. S. Park、Y. Zhang、Y. Jia、W. Han、C.-C. Chiu、B. Li、Y. Wu 和 Q. V. Le，《改进的自动语音识别噪声学生训练》，*arXiv* 预印本 *arXiv:2005.09629*，第 1–5 页，2020年。[65] G. Patel、J. P. Allebach 和 Q. Qiu，《Seq-ups：用于半监督文本识别的序贯不确定性意识伪标签选择》，收录于*IEEE/CVF* 冬季计算机视觉应用会议（*WACV*）论文集，2023年，第 6180–6190 页。[66] M. Barraco、M. Cornia、S. Cascianelli、L. Baraldi 和 R. Cucchiara，《CLIP 特征用于图像描述的惊人效能：实验分析》，收录于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集研讨会，2022年，第 4662–4670 页。[67] M. Tang、Z. Wang、Z. Liu、F. Rao、D. Li 和 X. Li，《Clip4caption：用于视频描述的 CLIP》，收录于*ACM* 国际多媒体会议（*ACM MM*）论文集，2021年，第 4858–4862 页。[68] C. Ma、Y. Yang、Y. Wang、Y. Zhang 和 W. Xie，《利用冻结视觉语言模型进行开放词汇语义分割》，*arXiv* 预印本 *arXiv:2210.15138*，第 1–21 页，2022年。[69] J. Ding、N. Xue、G.-S. Xia 和 D. Dai，《解耦零样本语义分割》，收录于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2022年，第 11 583–11 592 页。[70] M. Xu、Z. Zhang、F. Wei、Y. Lin、Y. Cao、H. Hu 和 X. Bai，《基于预训练视觉语言模型的零样本语义分割简易基线》，*arXiv* 预印本 *arXiv:2112.14757*，2021年。[71] C. Zhou、C. C. Loy 和 B. Dai，《从 CLIP 中提取免费密集标签》，收录于欧洲计算机视觉会议（*ECCV*）论文集，2022年，第 696–712 页。[72] Y. Rao、W. Zhao、G. Chen、Y. Tang、Z. Zhu、G. Huang、J. Zhou 和 J. Lu，《DenseCLIP：通过上下文感知提示实现语言引导的密集预测》，收录于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2022年，第 18 082–18 091 页。[73] Z. Wang、Y. Lu、Q. Li、X. Tao、Y. Guo、M. Gong 和 T. Liu，《CRIS：由 CLIP 驱动的指代表达图像分割》，收录于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2022年，第 11 686–11 695 页。[74] Z. Zhou、B. Zhang、Y. Lei、L. Liu 和 Y. Liu，《ZegCLIP：探索将 CLIP 适配于零样本语义分割》，*arXiv* 预印本 *arXiv:2212.03588*，2022年。[75] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，《Attention is All You Need》，收录于神经信息处理系统进展（*NeurIPS*），2017年，第 1–11 页。[76] J. F. Gemmeke、D. P. Ellis、D. Freedman、A. Jansen、W. Lawrence、R. C. Moore、M. Plakal 和 M. Ritter，《Audio Set：音频事件本体与人工标注数据集》，收录于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2017年，第 776–780 页。[77] K. He、X. Zhang、S. Ren 和 J. Sun，《用于图像识别的深度残差学习》，收录于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2016年，第 770–778 页。[78] D. Tran、H. Wang、L. Torresani、J. Ray、Y. LeCun 和 M. Paluri，《深入探究用于动作识别的时空卷积》，收录于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2018年，第 6450–6459 页。[79] A. Paszke、S. Gross、F. Massa、A. Lerer、J. Bradbury、G. Chanan、T. Killeen、Z. Lin、N. Gimelshein、L. Antiga 等，《PyTorch：一种命令式风格的高性能深度学习库》，神经信息处理系统进展（*NeurIPS*），2019年。