arXiv:2602.06765v1 [cs.SD] 6 Feb 2026

长音频中的层次化活动识别与字幕生成

Peng Zhang、*Qingyu Luo*、*Philip J.B. Jackson*、*Wenwu Wang*

中心 视觉、语音与信号处理中心（CVSSP），萨里大学，英国 .

摘要

现实世界音频中的复杂活动持续时间较长，并呈现层次化结构，但以往大多数研究聚焦于短片段和孤立事件。为弥补这一空白，我们推出*MultiAct*，这是一个用于从长音频中对人类活动进行多层次结构化理解的新数据集和基准。MultiAct包含时长较长的厨房录音，并在三个语义层级（活动、子活动和事件）上进行标注，同时配有细粒度描述和高层摘要。我们还提出了一种统一的层次化模型，可联合执行分类、检测、序列预测和多分辨率描述生成。在MultiAct上的实验建立了强有力的基线，并揭示了对长音频的层次化和组合结构进行建模时面临的关键挑战。未来一个值得探索的方向，是研究更适合捕捉长音频中复杂长程关系的方法。

索引词— 长音频、层次化活动识别、音频字幕生成、声音事件、MultiAct

1. 引言

尽管近年来音频理解日益受到关注，但仅凭听觉信号识别复杂人类活动的研究仍明显不足 [1, 2, 3]。尽管声音事件检测与字幕生成已取得近期进展，但现有的大多数数据集 [4, 5, 6]和模型 [7, 8, 9] 仍局限于短片段（例如10–30秒），通常只能捕捉孤立的声学事件或简短的动作片段。因此，对长时、结构化人类活动的理解在很大程度上仍有待探索。然而，在真实场景中，有意义的活动往往持续较长时间，并由时间上相关的事件序列构成 [10]。有效建模这种长时声学结构，尤其是事件之间的程序性关系，对于更深入的语义理解至关重要，但当前研究尚未充分解决这一问题 [11, 12]。

为了推进对长时活动的理解，关键在于对人类行为的结构复杂性进行建模，同时应对现有音频数据集的局限性 [13]。受生态心理学框架的启发 [14, 15], , 人类活动可以从三个级别进行分析：行为场景、行为片段和原子物理单元。行为场景指由特定空间、时间和社会约束界定的稳定活动模式。在厨房场景中，像清理这样的高级活动构成一种行为场景，用以组织目标导向的行为片段，其中包括擦洗平底锅等子活动。这些片段还可进一步分解为细粒度的原子物理单元，在听觉域中表现为流水声或金属碰撞等原始声音事件。因此，对这种层次结构进行建模，对于复杂听觉序列的稳健语义表示和推理至关重要。然而，AudioSet [4], 、VGGSound [5]和Clotho [6], 等广泛使用的音频数据集并不适合这一目的。如表1所示，这些数据集由带有扁平事件级标注的短片段构成，缺少研究组合语义和高级活动所需的时间范围、序列关系及层次组织。

表1。按标注层级比较音频数据集。

数据集	规模	平均时长	事件	子活动	活动
AudioSet [4]	∼5800小时	10s	✓	×	×
VGGSound [5]	∼550小时	10s	✓	×	×
Clotho [6]	∼24h	15秒–30秒（字幕）	×	×	
EPIC-SOUNDS [3]	∼100小时	∼4.9秒	✓	×	×
MultiAct	∼9h	–	✓	✓	✓

为应对这一挑战，我们引入了*MultiAct*数据集，这是一个用于从长音频对人类活动进行多层次结构化理解的新基准。MultiAct部分源自EPIC-SOUNDS [3], ，我们在其中对精选子集重新标注，并将其重组为带有层次时间标签和配对文本描述的长连续片段。MultiAct数据集具有三个显著特征：（1）三个语义级别的层次标注：高层活动、子活动和事件，从而能够研究组合结构与序列推理；（2）最长可达50分钟的长音频录音，使模型能够捕捉扩展的时间依赖与活动流程；（3）每个片段配有双重自然语言描述，包括细粒度描述和高层摘要，支持多分辨率语言生成。据我们所知，MultiAct是首个具备这些特征的音频数据集。

Furt此外，我们提出了一种统一的层次化框 ame-

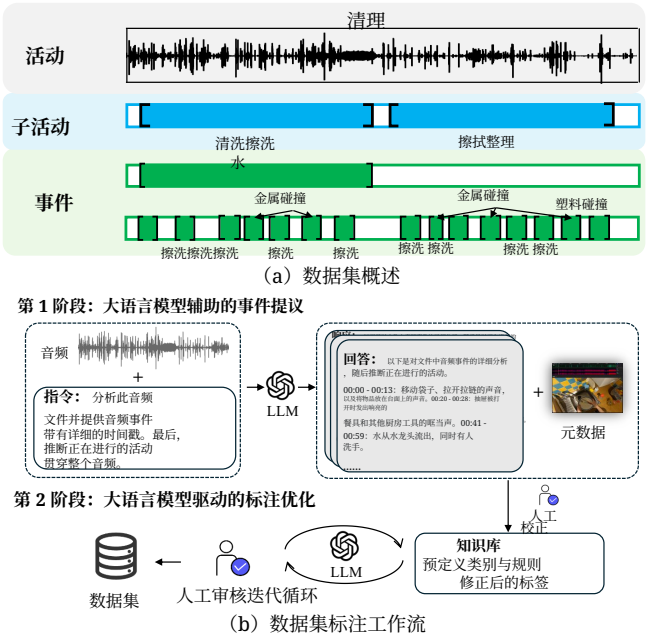


图1。*MultiAct* 的层次化结构及大语言模型辅助的标注流程。

架，用于理解长时音频活动。我们探索了层次化活动识别、子活动序列预测和多种分辨率下的字幕生成等任务。在MultiAct数据集上的实验表明，该方法在各项任务中均表现出色，并揭示了对结构化长时音频进行建模所面临的关键挑战。所有资源均可在以下网址获取：

`github.com/PennyZhang9/MultiAct`。

2. 数据集

MultiAct数据集源自个人在厨房环境中进行烹饪活动时的录音。该数据集基于活动部分学构建了三级层次结构，在活动、子活动和事件层级提供丰富的时间标注。如图1(a)所示，每个高级活动（例如，清理）由多个子活动（例如，清洗擦洗、擦拭整理）组成，而这些子活动又由一系列细粒度的声音事件构成（例如，擦洗、金属碰撞、水）。这种层次化结构使得我们能够在不同时间分辨率下对程序语义进行建模。

MultiAct 包含来自 17 个厨房环境的 8.97 小时长音频，这些音频取自 EPIC-SOUNDS [3]，并补充了额外的层次标注。它涵盖 3 个粗粒度活动类别中的 51 个活动实例，持续时间为 15.7 秒至 3164 秒（平均值为 628 秒）。在子活动级别，数据集包含 472 个经过标注的实例，涵盖 12 个类别，持续时间为 0.57 秒至 572 秒（平均值为 63.6 秒）。对于事件级建模，该分类体系沿用并细化了 44 个类别，共包含 7312 个实例，持续时间为 0.2 秒至 100 秒（平均值为 4.91 秒）。除了时间标注外，MultiAct 还包含与短时声学事件对齐的细粒度字幕和描述整体活动的高层摘要，支持在多个粒度级别上执行文本生成任务。MultiAct 的标注依据知识共享署名许可（CC BY）发布。配套音频取自 EPIC-SOUNDS [3]，，后者是 EPIC-KITCHENS [10]，的衍生项目，且仍受原 EPIC-KITCHENS 非商业研究许可约束。

如图1(b)所示，标注通过一个分两个阶段进行的大语言模型（LLM）辅助工作流程构建。在第 1 阶段，我们向 GPT-4o [16]，提供长时音频内容和特定于任务的指令，以生成结构化标注草稿。尽管由于音频复杂且时长较长，输出可能包含幻觉或边界错误，但它为后续细化提供了丰富的初始内容。在第 2 阶段，人工标注员在元数据以及预定义类别和标注规则知识库的指导下，对这些草稿进行迭代修正。鉴于该数据集采用多层级标注方案，这一人在回路流程尤为关键，并能确保所有语义层级的标签都高质量且一致。

3. 方法

如图 2 所示，我们提出了一个统一的层次化框架，用于长音频活动识别和说明文字生成。该框架由三个组件构成：听觉SlowFast（ASF） [17] 特征提取器、多层次层级活动编码器，以及用于生成说明文字和摘要的统一语言解码器。

采用同步快慢路径的 SlowFast 式编码器提取时间对齐的声学特征：快路径以高分辨率捕捉瞬态线索；慢路径则聚合较长时间范围内的频谱结构。融合后的输出构成供所有下游任务使用的词元化帧级表示。

为对多个时间分辨率下的语义结构进行建模，我们设计了一个由三个任务专用组件构成的层级编码器。**事件编码器** 执行细粒度事件分类和边界检测。**子活动编码器** 通过子活动分类、边界检测和序列预测来建模中层时间结构，并将其嵌入与帧级 ASF 特征融合，以进行活动级预测。**活动编码器** 预测高层过程类别。三个编码器均基于从 ASF 主干网络提取的共享音频词元表示运行，并分别进行训练，以生成特定级别的输出。

为生成自然语言输出，我们采用统一的基于BART的字幕生成解码器 [18]。ASF编码器提取的音频特征首先经过线性投影，再由多层Transformer音频编码器处理。还可以选择将层级编码器任一级别的语义线索（例如子活动序列或活动标签）输入预训练的BART文本编码器进行编码，以提供文本条件。随后，将音频编码器和文本编码器的输出拼接起来并输入BART解码器，由其自回归地生成

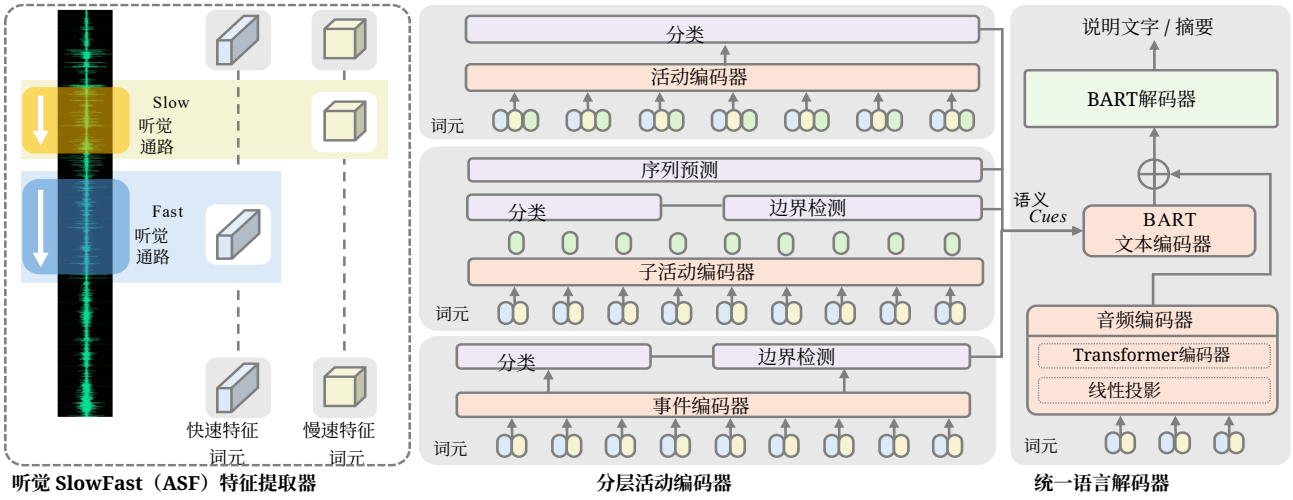


图 2。所提出架构的概览 .

细粒度字幕或高层摘要，具体取决于任务。

4. 实验

本节中，我们使用 MultiAct数据集在四项任务上评估我们的方法：**(i) 层次化活动分类**，给定时间边界明确的音频片段，在每个级别（活动、子活动、事件）预测一个标签；**(ii)音频活动检测**，在未剪辑音频上生成带类别标签的时间片段；**(iii) 音频活动序列预测**，预测有序的子活动序列，以从音频片段中捕捉时间动态和意图；**(iv) 音频活动描述**，为长时音频生成细粒度、按时间排序的说明文字或高层摘要。在所有实验中，基于 EPIC-SOUNDS [3] 预训练的 ASF主干网络 [17]用作冻结的特征提取器，生成 2304 维的帧级音频特征，并将其输入特定任务模型。

4.1. 分层活动分类

活动分类任务的目的是评估模型能否基于听觉信息，在多个语义层级上全面理解活动。给定一个具有时间边界的音频片段（含起止时间），模型需在活动、子活动和事件级别上各预测一个标签。该任务被形式化为三个并行的单标签多类分类任务。

基线。我们实现了三种基线，以研究架构选择。事件级分类通过在 ASF 音频特征之上训练一个线性层来完成。对于子活动和活动分类，我们冻结 ASF，并在其之上堆叠两个双向门控循环单元（BiGRU）层（每个方向包含 256 个隐藏单元）以编码时间动态，随后使用任务特定的注意力机制。ASF-Atten 对 BiGRU 输出应用自注意力池化 [19]以进行子活动分类。ASF-CrossAtten 引入交叉注意力模块 [20]，并联合训练子活动和活动分类头，以捕捉层级依赖关系。我们报告

表2。分类基线模型的结果（%，↑）。

划分	级别	模型	Top-1	Top-5	mPCA	mAP	mAUC
Val	事件	ASF	69.0	93.6	30.1	41.6	93.5
	子活动	ASF-Atten	50.0	**92.6**	55.2	**64.9**	85.5
		ASF-CrossAtten	**51.9**	88.9	**49.7**	65.9	**87.5**
	活动	ASF-Atten	60.0	N/A	60.7	**79.4**	87.2
		ASF-CrossAtten	**66.7**	N/A	**61.9**	72.7	84.6
Eval	事件	ASF	67.2	92.4	33.1	41.5	91.9
	子活动	ASF-Atten	**51.3**	**83.3**	**47.4**	**41.9**	**77.5**
		ASF-CrossAtten	42.3	74.4	33.8	37.0	76.4
	活动	ASF-Atten	83.3	N/A	83.3	**94.4**	**95.8**
		ASF-CrossAtten	**83.3**	N/A	**83.3**	72.2	70.8

五项标准指标：Top-1 和 Top-5 准确率、平均类别准确率（mPCA）、平均精度均值（mAP）以及曲线下面积均值（mAUC）。对于所有指标，数值越高表示性能越好。对于活动（3 个类别），Top-5 未定义，因此予以省略（单元格标记为不适用）。

结果。 表 2 展示了在验证集和评估集上三个语义级别的表现。ASF-CrossAtten 在活动分类上表现最佳（评估集 83.3%），而 ASF-Atten 在子活动方面表现出色，这得益于时序建模。

4.2. 音频活动检测

在本实验中，我们进一步评估时序检测任务。该任务旨在沿时间轴检测活动片段，并为每个片段分配一个类别标签（例如，子活动或事件）。这被表述为一个多类别时序检测问题，需要同时预测片段边界（起点和终点）以及每个检测到的片段所属的类别。

基线。我们将强大的视频动作检测模型 ActionFormer [21]，迁移至音频领域。训练期间，我们使用 2 秒窗口和 200 毫秒步长提取 ASF 特征，并将其输入 ActionFormer，以进行时间提案生成和分类。性能以多个交并比（IoU）阈值下的 mAP 报告。

表 3。检测基线模型的结果（%, ↑）。

划分	级别	不同交并比阈值下的平均精度					
		@0.1	@0.2	@0.3	@0.4	@0.5	Mean
Val	事件	17.0	14.7	12.8	11.2	9.8	13.1
	子活动	44.3	41.0	30.8	25.3	24.3	33.1
Eval	事件	16.5	15.7	15.0	13.6	12.5	14.6
	子活动	41.8	37.3	32.2	26.0	22.0	31.9

表 4。子活动序列预测模型的结果（%, ↓）。

划分	不同训练上下文长度下的 AER					
	2	3	4	6	8	Full
Val	66.7	72.2	75.9	81.5	88.9	79.6
Eval	69.2	74.4	75.6	87.2	87.2	80.8

olds.

结果。如表 3 所示，子活动检测在两个划分上的 AP 均持续优于事件检测。然而，在较高的交并比阈值下，性能明显下降。例如，在评估集上，子活动的 AP 从交并比为 0.1 时的 41.8% 降至交并比为 0.5 时的 22.0%，这表明精确定位边界仍面临挑战。

4.3. 音频活动序列预测

在音频数据集中，既有研究很少定义或系统性地研究活动序列预测。借助 MultiAct 密集且时间对齐的标注，以及丰富的子活动分类体系，我们将该任务形式化为序列预测：给定一个有界音频片段，预测按顺序排列的子活动序列。模型在没有明确片段边界的情况下，推断出现了哪些子活动及其顺序，重点关注流程结构和长程依赖，而非局部边界检测。

基线。 我们使用 ASF 音频特征作为 Conformer 编码器 [22] 的输入。首先通过线性层对这些特征进行投影。沿用原研究中大小为 31 的卷积核，我们将编码器配置为 8 层，并通过处理这些特征生成丰富的上下文表示。最后，线性层将这些表示映射为标签逻辑值，模型则使用连接时序分类（CTC）目标函数 [23] 进行训练，以预测子活动序列。在推理阶段，我们采用滑动窗口解码来处理长音频。我们评估活动错误率（AER）随训练上下文长度的变化情况，上下文长度分别设为 2、3、4、6 和 8 个子活动，并纳入整个序列（“ 完整 ”）进行比较。

结果。 表 4 显示，较短的上下文（2–4）产生的 AER 最低，而较长的上下文（6–8）会导致错误增加，使用 Full 时也仅有部分改善。这些结果凸显了局部上下文的重要性，以及对长程依赖进行建模的难度，说明序列预测仍是一项具有挑战性的开放问题。

表 5。细粒度字幕生成（C）和高级摘要（S）基线模型的结果（%, ↑）。

划分	Task	方法	BLEU1	BLEU4	METEOR	ROUGE-L	CIDEr
Val	C	Rule	24.0	4.9	11.4	20.0	3.8
		层次化	**39.2**	**10.8**	**15.8**	**28.5**	**16.2**
	S	Rule	23.7	7.99	11.7	26.9	9.3
		层次化	**28.4**	**9.8**	**13.5**	**32.5**	**24.1**
Eval	C	Rule	16.8	2.7	**13.0**	20.2	2.8
		层次化	**17.8**	**3.4**	9.6	**23.0**	**20.1**
	S	Rule	17.3	**7.5**	8.8	20.7	2.2
		层次化	**21.8**	7.3	**11.0**	**28.3**	**11.1**

4.4. 音频活动描述

在本实验中，我们探究显式建模事件、子活动与活动之间的层次关系，能否提升长音频字幕的连贯性与语义保真度。该任务侧重于从两个级别生成长音频片段的自然语言描述：细粒度描述和高层摘要。

基线。 为建立字幕生成基线，我们采用双编码器、BART 解码器 [18] 架构。ASF派生音频特征经线性投影至 BART 隐藏层维度，并由 Transformer音频编码器（10 层、8 个注意力头）进行编码。与此同时，BART文本编码器处理来自多层次活动编码器的结构化线索。随后，我们将音频编码器和文本编码器的输出拼接起来，并输入 BART解码器生成字幕。训练采用教师强制和词元级交叉熵；推理时采用束搜索（束宽为 4），并结合长度惩罚和三元组阻断。我们使用 BLEU-1/4 [24],METEOR [25], ROUGE-L [26],和 CIDEr [27]评估性能。

结果。 表 5 报告了验证集和评估集上细粒度描述（C）与高层摘要（S）的结果。层次化模型在大多数指标上优于基于规则的基线，其中 CIDEr 和 ROUGE-L 的提升最大。与字幕生成相比，摘要通常取得更高的 BLEU、METEOR 和 ROUGE 分数，而字幕生成在评估集上的 CIDEr 分数最高。总体而言，摘要的表现优于细粒度描述，这凸显了为长音频生成时间精确且语义丰富的描述所面临的挑战。

5. 结论

我们推出 MultiAct，这是首个面向结构化长音频的基准，提供层次标注（活动、子活动、事件），并配有细粒度字幕和高层摘要。除数据集外，我们还提出一种用于联合识别与描述生成的统一层次模型。在 MultiAct 验证集上的实验验证了该方法，并凸显了长音频语义与流程结构建模中持续存在的挑战。我们希望这项工作能推动长音频高层活动理解持续进步。

6. 致谢

感谢 Pablo Mart´ınez-Nuevo、Sven Ewan Shepstone 和 Jon Francombe（Bang & Olufsen A/S）提供宝贵的讨论意见。本研究得到 Bang & Olufsen A/S 的支持。

7. 参考文献

[1] J. Shin、N. Hassan、A. S. M. Miah 和 S. Nishimura，《跨多种数据模态的人类活动识别综合方法综述》，《*Sensors*》，第 25 卷，第 13 期，第 4028 页，2025年。[2] G. Wijngaard、E. Formisano、M. Esposito 和 M. Dumontier，《场景与事件的音频-语言数据集：综述》，《*IEEE Access*》，2025年。[3] J. Huh、J. Chalk、E. Kazakos、D. Damen 和 A. Zisserman，《EPIC-SOUNDS：大规模声音动作数据集》，《*IEEE TPAMI*》，2025年。[4]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》，载于*ICASSP*，2017年，第 776–780 页。[5] H. Chen、W. Xie、A. Vedaldi 和 A. Zisserman，《VGGSound：大规模视听数据集》，载于*ICASSP*，2020年，第 721–725 页。[6] K. Drossos、S. Lipping 和 T. Virtanen，《Clotho：音频描述数据集》，载于*ICASSP*，2020年，第 736–740 页。[7] Y. Hou、B. Kang、A. Mitchell、W. Wang、J. Kang 和 D. Botteldooren，《用于声学场景分类的协同场景-事件建模》，《*IEEE/ACM TASLP*》，第 32 卷，第 82 页，2023年，第 68 页。[8] X. Mei、C. Meng、H. Liu、Q. Kong、T. Ko、C. Zhao、M. D. Plumbley、Y. Zou 和 W. Wang，《WavCaps：用于音频-语言多模态研究的ChatGPT辅助弱标注音频描述数据集》，《*IEEE/ACM TASLP*》，第 32 卷，第 3339–3354 页，2024年。[9] J. Hu、Y. Cao、M. Wu、F. Kang、F. Yang、W. Wang、M. D. Plumbley 和 J. Yang，《PSELDNets：基于大规模合成数据集预训练的声音事件定位与检测神经网络》，《*IEEE/ACM TASLP*》，2025年。[10] D. Damen、H. Doughty、G. M. Farinella、S. Fidler、A. Furnari、E. Kazakos、D. Moltisanti、J. Munro、T. Perrett、W. Price 等，《EPIC-KITCHENS 数据集：收集、挑战与基线》，《*IEEE TPAMI*》，第 43 卷，第 4141 页，2020年，第 11 期，第 4125 页。[11] Q. Kong、Y. Cao、T. Iqbal、Y. Wang、W. Wang 和 M. D. Plumbley，《PANNs：用于音频模式识别的大规模预训练音频神经网络》，《*IEEE/ACM TASLP*》，第 28 卷，第 2880–2894 页，2020年。[12] Y. Hou、S. Song、C. Yu、W. Wang 和 D. Botteldooren，《用于声学场景分类的音频事件关系图表示学习》，《*IEEE Signal Processing Letters*》，第 30 卷，第 1382–1386 页，2023年。[13] X. Mei、X. Liu、M. D. Plumbley 和 W. Wang，《自动音频描述：近期进展概述与新挑战》，《*EURASIP Journal on Audio, Speech, and Music Processing*》，第 2022 卷，第 1 期，第 26 页，2022年。[14] R. G Barker 和 H. F Wright，《美国中西部及其儿童：一个美国城镇的心理生态学》，*Row, Peterson andCompany*，1955年。[15] Z. Luo、W. Xie、S. Kapoor、Y. Liang、M. Cooper、J. C. Niebles、E. Adeli 和 F.-F. Li，《MOMA：多目标、多参与者活动解析》，神经信息处理系统进展，第 34 卷，第 17 939–17 955 页，2021年。[16] A. Hurst、A. Lerer、A. P. Goucher、A. Perelman、A. Ramesh、A. Clark、A. Ostrow、A. Welihinda、A. Hayes、A. Radford 等，《GPT-4*21276*，2024*2410*o 系统卡》，*arXiv* 预印本 *arXiv:* 。[17] E. Kazakos、A. Nagrani、A. Zisserman 和 D. Damen，《用于音频识别的慢速-快速听觉流》，载于*ICASSP*，2021年，第 855–859 页。[18] M. Lewis、Y. Liu、N. Goyal、M. Ghazvininejad、A. Mohamed、O. Levy、V. Stoyanov 和 L. Zettlemoyer，《BART：面向自然语言生成、翻译和理解的序列到序列去噪预训练》，载于*ACL*，2020年，第 7871–7880 页。[19] Y. Zhu、T. Ko、D. Snyder、B. Mak 和 D. Povey，《用于文本无关说话人验证的自注意力说话人嵌入》，载于*Interspeech*，2018年，第 2–6 页。[20] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. Kaiser 和 I. Polosukhin，《注意力就是一切》，神经信息处理系统进展，第 30 卷，2017年。[21] C.-L. Zhang、J. Wu 和 Y. Li，《ActionFormer：使用Transformer定位动作片段》，载于*ECCV*，2022年，第 492–510 页。[22] A. Gulati、J. Qin、C.-C. Chiu、N. Parmar、Y. Zhang、J. Yu、W. Han、S. Wang、Z. Zhang、Y. Wu 等，《Conformer：用于语音识别的卷积增强型Transformer》，载于*Interspeech*，2020年，第 5036–5040 页。[23] A. Graves、S. Fern´andez、F. Gomez 和 J. Schmidhuber，《连接主义时序分类：使用循环神经网络标注未分段序列数据》，载于*ICML*，2006年，第 369–376 页。[24] K. Papineni、S. Roukos、T. Ward 和 W.-J. Zhu，《BLEU：一种机器翻译自动评估方法》，载于*ACL*，2002年，第 311–318 页。[25] S. Banerjee 和 A. Lavie，《METEOR：一种改进与人工判断相关性的机器翻译评估自动指标》，载于*ACL*机器翻译和/或摘要内在与外在评估指标研讨会论文集，2005年，第 65–72 页。[26] C.-Y. Lin，《ROUGE：摘要自动评估工具包》，载于文本摘要分支拓展：*ACL-04*研讨会论文集，2004年，第 74–81 页。[27]R. Vedantam、C. L. Zitnick 和 D. Parikh，《CIDEr：基于共识的图像描述评估》，载于*CVPR*，2015年，第 4566–4575 页。