arXiv:2603.10623v1 [eess.AS] 11 Mar 2026

Geo-ATBench：结合地理空间语义上下文的地理空间音频标注基准

Yuanbo Hou[a, 1, *], Yanru Wu[b, 1], Qiaoqiao Ren[c], Shengchen Li[b], Stephen Roberts[a], Dick Botteldooren[d]

[a]牛津大学工程科学系机器学习研究组，英国[b]西交利物浦大学智能科学系，中国[c]瑞典皇家理工学院电气工程与计算机科学系，瑞典[d]根特大学信息技术系*WAVES*研究组，比利时

摘要

计算听觉场景分析（CASA）中的环境声音理解通常被表述为一个仅音频识别问题。这种表述在多标签音频标注（AT）中存在一个长期存在的缺陷：声学相似性可能使某些事件仅凭波形难以区分。在此类情况下，消歧线索往往存在于波形之外。地理空间语义上下文（GSC）源自地理信息系统数据，例如兴趣点（POI），能够提供与位置相关的环境先验，有助于减少这种歧义。通过所提出的地理空间音频标注（Geo-AT）任务，可以系统地研究这一方向；该任务在音频之外结合GSC，对多标签声音事件标注进行条件化。为对Geo-AT进行基准测试，本文引入了Geo-ATBench数据集，作为一个带有地理标注的复音音频基准，包含涵盖28个事件类别的10.71小时真实世界音频；每个音频片段都配有一个基于兴趣点（POI）的GSC表征，该表征由11个语义上下文类别构建而成。此外，本文提出了GeoFusion-AT，作为统一的地理空间音频融合框架，用于

*通讯作者：Yuanbo Hou，牛津大学机器学习研究组，英国。电子邮箱：Yuanbo.Hou@eng.ox.ac.uk[1]贡献相同。

在三个具有代表性的音频骨干网络上评估特征级、表征级和决策级融合，并设置仅音频和仅GSC基线。实验表明，引入GSC通常能够提升AT性能，尤其是在声学混淆标签上，这说明地理空间语义能够提供一种超越单纯音频的有效先验。一项针对10名参与者和579个样本开展的众包听音研究显示，在Geo-ATBench标签与汇总人工标签上，各模型之间的性能均无显著差异，支持Geo-ATBench成为一个与人类判断一致的基准。总体而言，所提出的Geo-AT任务、开放基准Geo-ATBench以及可复现的地理空间音频融合框架GeoFusion-AT，为在CASA社区内研究结合地理空间语义上下文的音频标注提供了坚实基础。有关数据集、源代码和模型，请参阅项目主页（https://github.com/WuYanru2002/Geo-ATBench）。 关键词：
计算听觉场景分析、多标签音频标注、地理空间语义上下文、兴趣点、多模态融合

1. 引言

环境声音理解是计算听觉场景分析（CASA） [1]的核心目标之一。在许多实际应用中，目标输出是多标签音频标注（AT） [2], ，其中每段录音可能包含多个声音事件，系统预测事件标签集合。AT支持声学监控 [3], 、智慧城市感知 [4], 、多媒体检索 [5], 以及智能家居助手 [6]等应用。

尽管用于环境音频的深度学习模型取得了显著进展，音频标注（AT）通常仍被视为纯音频识别问题 [7, 8]。近期的音频标注骨干网络，包括卷积神经网络（CNNs）和 Transformer，能够从时频特征中学习强大的声学表示

例如梅尔频谱图 [9]。然而，一个长期存在的缺陷仍未解决：仅凭波形，声学相似性可能会使某些事件难以区分，尤其是在不同声源产生高度相似的时频模式时 [10, 11]。在这种情况下，消除歧义的线索往往不在波形之中 [12, 13]。这类线索的一个重要来源是声音发生时所处的物理环境。声音事件由嵌入特定地点的声源产生，其发生受到与位置相关的环境因素影响 [14]。与位置相关的条件会使事件标签与地理空间语义上下文（GSC）之间形成系统性关联 [15]。因此，当仅凭波形无法消除歧义时，GSC 能够提供互补线索。

本研究聚焦于与声源相关的GSC，即从地理信息系统数据（如兴趣点（POI） [16]）中获取、与位置相关联的环境先验。与原始GPS坐标相比，基于POI的GSC能够对声源周围的物理环境进行结构化语义描述，并可与音频表征对齐 [17]。由于缺乏标准化任务，以及在可复现评估条件下将音频与可靠、结构化GSC配对的基准数据集 [18]，这一方向的进展仍受到限制。近年来，移动录音设备和位置感知媒体平台日益将录音与地理坐标关联起来 [18], ，使相关的音频-GSC配对数据越来越容易获取。这一趋势为研究如何利用GSC支持真实世界中的多标签音频标注创造了适逢其时的契机。

为弥补音频标注通常未结合与声源位置相关联的 GSC 这一不足，本文提出地理空间音频标注（Geo-AT）任务，该任务在音频录音之外，还以 GSC 为多标签音频标注提供条件。Geo-AT 旨在评估位置相关的环境先验是否有助于消歧那些仅凭音频难以区分的事件。为对 Geo-AT 进行基准测试，我们发布了 Geo-ATBench 数据集，这是一个经过地理标注的复音音频基准，包含 3,854 个音频片段和 28 个事件标签；每个音频片段都配有一个基于 11 种情境类别中的兴趣点语义构建的 GSC 表征，从而能够对地理空间语义如何与多标签音频标注中的声学表示相互作用开展可复现研究。

cally标注的复音音频基准，包含3,854个音频片段及28个事件标签；每个音频片段均配对一个由11个情境类别上的兴趣点语义构建的GSC表征，从而支持对地理空间语义如何与多标签音频标注中的声学表示相互作用进行可复现研究。

所提出的 Geo-ATBench 基准设计并未规定应如何将 GSC 集成到音频标注模型中 [2, 3, 19], ，不同的集成选择可能导致不同的结果。因此，本文引入 GeoFusion-AT，作为面向所提出 Geo-AT 任务的统一地理空间音频融合框架，用以对代表性融合策略进行基准评测，并在 Geo-ATBench 上报告参考结果。具体而言，GeoFusion-AT 在三种代表性音频骨干网络上评估三种典型融合策略，即特征级、表示级和决策级融合；这三种骨干网络分别是基于 CNN 的预训练音频神经网络（PANNs） [20], 、基于 Transformer 的音频频谱图 Transformer（AST）[9], 以及对比式语言-音频预训练（CLAP） [21]。同时纳入仅音频基线与仅 GSC 基线，以分离各模态的贡献，并识别融合在何种情况下能够超越任一单一输入的性能。

主要贡献如下：1）提出 Geo-AT，作为 CASA 中多标签音频标注的标准化任务定义，将音频与地理空间语义上下文（GSC）相结合；2）发布 Geo-ATBench，作为可复现 Geo-AT 评估的开放基准，包含 3,854 个带有 28 个事件标签的真实复音音频片段；每个音频片段均配有基于兴趣点语义构建的 GSC 表征，涵盖 11 个语义上下文类别；3）提出 GeoFusion-AT，这是一种统一的地理空间音频融合框架，在 Geo-ATBench 上针对具有代表性的音频骨干网络对代表性融合策略进行基准测试，并报告参考结果；4）邀请 10 名参与者对 579 个样本开展众包听音研究。研究表明，以 Geo-ATBench 标签和汇总人工标签作为评估依据时，模型性能具有可比性，从而支持将 Geo-ATBench 作为一个

与人类判断一致的基准。我们已发布数据集、代码和模型。

本文其余部分安排如下。第2节回顾相关工作，以明确 Geo-AT 的定位。第3节对 Geo-AT 任务进行形式化定义。第4节介绍 Geo-ATBench数据集。第5节介绍 GeoFusion-AT 框架，以及基于代表性音频骨干网络的融合策略。第6节报告实验结果并进行分析。第7节详述人工评估研究。第8节总结全文。

2. 相关工作

本节将提出的地理空间音频标注（Geo-AT）任务置于多标签音频标注（AT）、上下文感知声音理解，以及从地理信息系统中获取的基于POI的地理空间语义上下文等相关研究中进行定位。相关讨论阐明了在可复现评估下建立标准化 Geo-AT 任务的必要性。

2.1. 多标签音频标注与声学歧义

多标签音频标注是 CASA 中的一项核心任务 [1], ：一个音频片段可能包含复音声音事件，目标是预测其事件标签集合 [2]。大规模基准测试和挑战赛 [22, 23] 推动了模型架构与骨干网络的持续发展，例如基于卷积神经网络的 PANNs [20]和 MobileNet [24], 、基于 Transformer 的分层词元语义音频 Transformer [25] 和 AST [9], ，以及基于对比学习的 CLAP，它能够对齐音频与语言表征 [21]。这些骨干网络已成为音频标注任务中表征学习的常用参照。

尽管架构不断进步，真实环境中的音频标注仍持续面临难以消除的歧义 [26]。复音录音常包含相互重叠的声源，不同事件也可能产生相似的时频模式 [10, 11], ，导致仅音频分类难以区分易混淆事件，并出现误

分类。与声源相关的GSC等外部先验，通过将与位置相关的环境先验编码为结构化的基于POI的语义表示 [16], ，提供互补线索，例如附近的地点类别及其围绕声源的分布。与位置相关的GSC可约束场景中可能发生的事件范围，并在单靠声学证据不足时帮助消歧。

2.2. 声音理解的上下文与辅助信息

上下文感知的声音理解通过纳入声学表示之外的信息，扩展了音频标注。先前的研究 [27, 28] 可分为两类，区别在于附加信号是否与音频在时间上对齐。一类研究 [29] 使用配对的感知流，其中视频帧或其他时间对齐的输入与音频同时提供。另一类研究 [30] 使用与录音环境相关联、但不与音频信号在时间上对齐的辅助元数据。

Geo-AT关注第二类。与位置相关的描述符可作为场景级先验，在许多部署场景中均可获取。现有研究 [31]在辅助元数据的纳入方式上各不相同，涉及元数据表示、音频-元数据配对规则、数据划分和报告惯例。研究并不总是报告仅音频和仅元数据基线。这些不一致限制了不同研究之间可复现的比较，也促使人们建立一项标准化任务，以评估辅助元数据在多标签音频标注中的作用。

2.3. 基于*POI*的地理空间语义上下文（*GSC*）

由于移动录音设备和位置感知媒体平台会将录音与地理坐标关联起来，音频采集中可获取的地理空间信息日益增多 [18]。一些数据集包含地理或位置相关标注，从而能够对城市

声音环境和区域差异进行空间分析 [15]。然而，地理空间信息通常用于组织、制图或描述性分析，而不是作为声音事件识别的显式模型输入 [32]。

地理信息系统中的兴趣点（POI）将位置转化为可解释的语义描述符。POI编码附近地点及其构成，表示与位置相关的环境先验 [16]。基于POI的GSC包含可与音频录音配对的场景级描述符。然而，先前工作很少将基于POI的GSC形式化为音频标注任务的一部分。缺乏一致的任务定义和基准测试，使得难以评估是否以及如何整合地理空间语义。

*2.4. Geo-AT*的定位

综合来看，先前工作使音频标注在很大程度上保持仅音频，并且很少在可复现协议下将基于POI的GSC作为任务输入进行评估。缺失的部分是标准化地理空间音频标注任务定义，以及一个能够实现受控比较的基准测试。
Geo-AT 任务通过将音频标注定义为以与声源相关的、基于POI的GSC以及音频为条件，填补了这一空白，从而能够对音频标注任务中的地理空间先验进行受控评估。

3. 所提出的地理空间音频标注（Geo-AT）任务

地理空间音频标注（Geo-AT）将音频标注形式化为以声源相关的地理空间语义上下文（GSC）为条件，该上下文源自地理信息系统资源，例如兴趣点（POI）。Geo-AT是一项多模态学习任务，能够对基于POI的GSC如何与音频标注任务中的声学表示相互作用进行受控研究 [2, 19]。

3.1. 问题定义

鉴于每段录音都由声学表征 $\mathbf{A}$ 和一个由地理信息系统构建的 GSC向量 $\mathbf{g} \in \mathbb{R}^{D_{\mathrm{GSC}}}$ 表示，Geo-

AT 使用成对输入（$\mathbf{A}$，$\mathbf{g}$）。Geo-AT 假设在推理时可将 $\mathbf{g}$ 作为录音元数据，与 $\mathbf{A}$一同提供。学习目标是预测音频片段中出现的事件标签集合。令 $\mathcal{Y}$ 表示事件标签集合。每个音频片段的目标是一个多标签向量 $\mathbf{y} \in \{0,1\}^{|\mathcal{Y}|}$，其中 $\mathbf{y}_k = 1$ 表示事件 k 是否出现在音频片段中。Geo-AT旨在学习一个函数 f：（$\mathbf{A}$，$\mathbf{g}$）$\rightarrow$$\mathbf{y}$，其中 $\mathbf{g}$ 通过基于兴趣点的语义描述符（例如，与海滩、高速公路、火车站、住宅区或工业设施的距离）来编码周边环境信息。Geo-AT并未规定 $\mathbf{A}$ 与$\mathbf{g}$之间的具体集成机制，因此在共同的任务定义下将模型设计选择留待评估。

3.2. 任务范围

Geo-AT侧重于多标签标注，而非单标签分类，强调在复音条件下进行标签预测，此时一个音频片段中可能同时出现多个事件。Geo-AT的目的并非取代音频标注任务，而是研究空间证据何时以及如何补充音频表征，尤其是对于声学上易混淆的事件和复音重叠。Geo-AT的研究动机源于听觉感知中上下文知识的运用，以及实际部署中的位置关联元数据 [33]。Geo-AT为构建和评估更稳健的机器听觉系统提供了框架，可应用于地理环境多样的场景，包括城市噪声监测、情境感知辅助听力和可扩展的声学监控 [18] [34]。

4. Geo-AT任务基准数据集：Geo-ATBench

4.1. 数据收集

Geo-ATBench 的音频录音取自 Freesound.org [35]，这是一个由用户贡献声音的公共存储库；此外，还取自 [33]，中介绍的数据集，该数据集包含带有全球定位系统信息的音频文件，涵盖多种声音

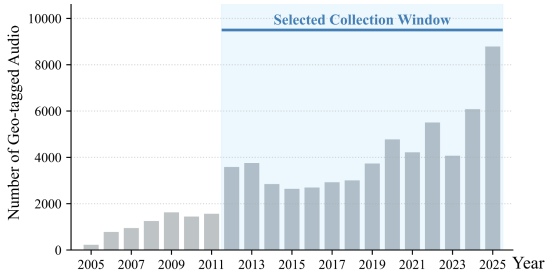


图 1：每年上传至 Freesound 且包含全球定位系统信息的录音数量。

事件。音频片段根据是否包含地理标注信息进行筛选，具体来说，即上传者提供的纬度和经度坐标；随后对坐标的有效性以及标签与位置之间明显不匹配的情况进行仔细的人工审核，以确保质量。

4.2. 声音事件和 *GSC* 标注

GSC 构建：对于来自 Freesound 的录音，我们特意选取了 2012 年至 2025 年间的数据。如图 1 所示，2012 年以前带有地理标签的音频数量相对有限，因此进行这一时间筛选；此外，随着时间跨度拉长，各地区的地理信息也可能发生变化。每段录音的 GPS坐标均取自 Freesound，其他录音的坐标则取自其原始数据集。我们利用这些坐标，通过 Overpass API [36] 查询 OpenStreetMap (OSM) 地理空间数据库。对于每段带有 GPS坐标的录音，我们以其位置为中心绘制一个固定边长的正方形，并依据 11 个 OSM要素键识别该正方形内的 OSM实体，涵盖土地利用、便利设施和自然等类别。圆形区域在概念上或许更符合声音传播的各向同性特征，但为便于在标准的基于 OSM 的地理信息系统中高效执行边界框查询，我们采用正方形区域。这一选择在保持各样本空间一致性的同时，为本地声学环境提供了计算上实用的近似。由此得到的 GSC 表征是一种从这些 OSM 标注中提取的、基于兴趣点的语义

类别	数量		数量		数量	
自然声音			人类声音		物体声音	
鸟类声音	1024	8191		794 5133		463 3068
蟋蟀	343	3091		288 2225		340 3092
落水声	325	2922		188 1593		165 1291
流水	319	2774		144 1330		121 835
波浪	307	2754		81 624		115 927
飞虫	137	824		79 249		111 731
Wind	83	737		53 125		107 293
					爆炸	93 431
					Bus	74 461
					警笛声	69 509
					地铁	63 454
					直升机	58 496
					Dog	209 969
					卡车	42 237

表1：Geo-ATBench 中的声音类别，按自然、人类和物体分组。时长表示每个类别的总时长（秒），数量表示样本数。乐器是 Musical instrument 的缩写，落水声表示落水声/雨。

描述符，用作第 3 节所述的位置关联输入。所有音频片段均使用相同的正方形边长和 11 个要素键，以确保整个数据集中的 GSC提取保持一致。

声音事件标注：许多 Freesound 音频片段都包含用户提供的标签，而音频事件的感知通常基于人类听觉。因此，我们会逐一手动审核每段录音，聆听音轨并为听到的事件分配标签。若标签存在不确定性，我们会反复播放并重新检查录音，直到能够作出判断。完成手动标注后，我们会将标签与 Freesound.org[35] 上用户提供的标签交叉验证。对于存在分歧的录音，我们会重新检查并更正；必要时，还会利用相应的全球定位系统元数据提取基于兴趣点的OSM注释，作为辅助线索来支持标签核验。审核完成后，每段音频片段都会与其基于兴趣点的OSM注释配对，形成 Geo-ATBench 中的音频–GSC对。初始标注耗时约 600 人时，交叉验证和复核又耗时约 200 人时，四个月内总计约 800 人时。从头到尾，数据集收集、整理和标注共耗时约六个月。

通过整理流程，将非结构化标注标签映射到

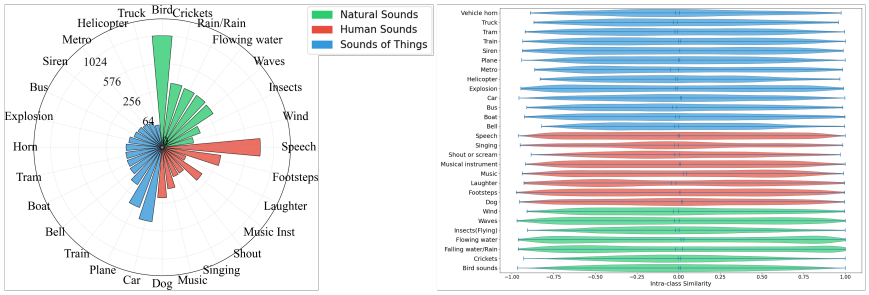


图 2：声音类别与声学相似性的概述。（左）三种粗粒度声音类别的分布。（右）基于对数梅尔频谱特征计算的 28 种声音事件类别的类内相似性。

受控词表中，最终得到28种声音事件类别。这28种类别被划分为三个主要类别，与AudioSet的分类结构 [26]相对应：1) **自然声音**，包括源自自然界的声音；2)**人类声音**，包括人类发出的声音；以及3) **物体声音**，代表机械和人造噪声。这些类别的样本数量和总时长见表1，而粗粒度分布及相应的类内相似度则显示在图2（右）中。小提琴图中的相似度根据对数梅尔频谱特征计算，特征向量之间的余弦相似度用于度量相似性，这是音频与声音分析中广泛采用的指标 [37]。此外，图3概述了数据集的构成，包括28种事件类型和11个OSM类别。该数据集本质上是多标签的，能够记录单个10秒录音中多个声音事件共同出现的情况。

4.3. 数据集组织与统计

经过清理和筛选，最终的 Geo-ATBench数据集包含 3,854 个音频片段，音频总时长为 10.71 小时。每个数据点由一个三元组构成：(i) 一个 10 秒的音频片段；(ii) 一个涵盖 28 个事件类别的片段级多标签向量；以及 (iii) 一个基于 OSM 标注构建的、涵盖 11 个语义上下文类别的基于兴趣点的GSC表示。为确保一致性

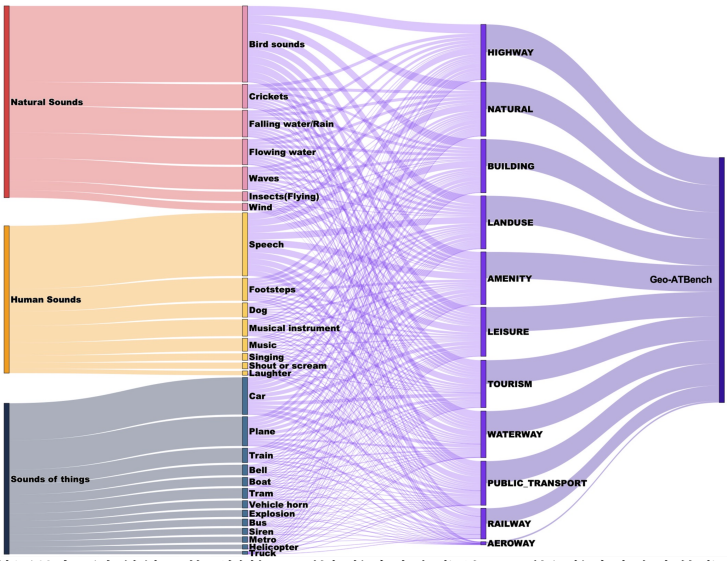


图 3：桑基图从左至右总结了共现链接：3 种粗粒度声音类别、28 种细粒度声音事件类别、GSC类型以及 Geo-ATBench数据集。流的宽度表示共现强度。该图展示了数据集中音频事件和 GSC类型的分布，并非旨在体现精确的现实世界关系，因为声音的出现可能因具体地理环境（例如住宅道路与高速公路）而存在显著差异。

以用于建模任务，所有收集到的录音均经过处理，转换为标准化格式。每个音频片段时长固定为 10 秒，编码为单声道 WAV文件，采样率为 16 kHz，位深度为 16 位。有关数据集的更多详情及获取方式，请访问项目主页。

5. GeoFusion-AT 框架及其实例化

5.1. GeoFusion-AT 框架

如图 4 所示，GeoFusion-AT 为 Geo-ATBench 数据集上的 Geo-AT 任务提供了三种典型融合点的参考实现。所有变体均采用成对输入（$\mathbf{A}$,$\mathbf{g}$），并输出事件类别的多标签logits $\mathbf{z} \in \mathbb{R}^C$ C，随后通过 Sigmoid函数计算标签概率。

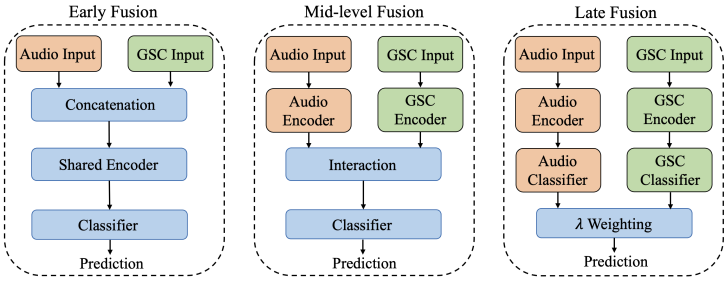


图 4：用于 Geo-AT 任务的 GeoFusion-AT 框架整体架构。

5.1.1. GeoFusion-Early：特征级融合

早期融合 [38], 也称为特征级融合，在网络输入端整合地理空间上下文和声学信息。该过程首先将原始音频波形转换为对数梅尔频谱图$\mathbf{A} \in \mathbb{R}^{1\times T\times F}$，其中 T和 F分别表示时间帧数和频率频点数。同时，GSC向量$\mathbf{g} \in \mathbb{R}^{D_{\text{GSC}}}$通过线性变换投影为长度为$F$ 的向量$\mathbf{g}' \in \mathbb{R}^{F}$ ： $\mathbf{g}' = \mathbf{W}_{\text{proj}}\mathbf{g}$，其中$\mathbf{W}_{\text{proj}} \in \mathbb{R}^{F\times D_{\text{GSC}}}$ 是可学习的投影矩阵。我们选择投影的频率分辨率 F ，使$\mathbf{g}'$ 可被视为位置条件化频谱先验（即逐频率加权/门控信号）： 不同地理环境往往与不同的主要声源和背景噪声相关联，并表现为各频带上具有特征性的能量分布。随后，投影后的向量$\mathbf{g}'$ 沿时间维度广播，形成广播 GSC 张量$\mathbf{G} \in \mathbb{R}^{1\times T\times F}$。

将音频频谱图和广播 GSC 张量沿通道维度拼接，得到融合表示 $\mathbf{X}_{\text{fused}} = \text{Concat}(\mathbf{A}, \mathbf{G}) \in \mathbb{R}^{2\times T\times F}$，并将其作为骨干网络的输入。当骨干网络不接受双通道频谱图输入时，会使用输入适配器将$\mathbf{X}_{\text{fused}}$ 映射为骨干网络所需的输入形状和通道格式；其余骨干组件均保持不变。

5.1.2. GeoFusion-Inter：表征级融合

中间融合 [39], 或表征级融合，是指在各模态分别经过独立编码器处理后，在潜在空间中组合信息。设 Φ_{audio} 为一个音频编码器，它将输入频谱图$\mathbf{A}$映射为音频嵌入$\mathbf{E}_{\text{audio}} \in \mathbb{R}^{D_{\text{emb}}}$，其中 D_{emb} 为嵌入维度。同样，GSC向量$\mathbf{g}$ 经过多层感知机（MLP）投影，生成维度相同的 GSC 嵌入$\mathbf{E}_{\text{GSC}} \in \mathbb{R}^{D_{\text{emb}}}$ 。此处，两个嵌入均为片段级表征，这意味着在融合之前，$\mathbf{A}$ 中的时间信息已由 Φ_{audio} 聚合。

中间融合采用对称跨模态注意力机制 [40]模块，支持音频嵌入与 GSC 嵌入之间的双向细化。设$\mathbf{Q}$，$\mathbf{K}$ 和 $\mathbf{V}$ 分别为查询、键和值，则注意力计算为 $\text{Attention}(\mathbf{Q},\ \mathbf{K},\ \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{D_{\text{emb}}}}\right)\mathbf{V}$，其中$\mathbf{K} = \mathbf{V}$，因子 $\sqrt{D_{\text{emb}}}$可稳定优化 [40]。因此，跨模态注意力作用于全局嵌入，而非时间标记，发挥特征级条件化的作用，~~而不是~~进行帧级对齐。将音频嵌入$\mathbf{E}_{\text{audio}}$作为查询，并将 GSC 嵌入$\mathbf{E}_{\text{GSC}}$ 作为键/值，可增强音频嵌入。对称地，使用$\mathbf{E}_{\text{audio}}$ 作为上下文来增强$\mathbf{E}_{\text{GSC}}$ 。通过残差混合形成两条互补的融合流。一条流将经过交叉注意力细化的音频嵌入与原始 GSC 嵌入结合。另一条流则将经过交叉注意力细化的 GSC 嵌入与原始音频嵌入结合。这种对称设计既保留了跨模态更新，也保留了原始模态信息。随后，将两条流拼接并通过可学习的线性投影，生成单个融合嵌入，再将其输入分类头，输出多标签标注逻辑值。

5.1.3. GeoFusion-Late：决策级融合

后期融合 [38], 或决策级融合，是指结合两个独立流的输出，每种模态各对应一个流。在这种范式下，音频分支Φ_{audio}处理音频表示$\mathbf{A}$ 并生成按类别划分的逻辑值$\mathbf{z}_{\text{audio}} \in \mathbb{R}^C$，其中 C 为事件类别的数量。与此同时，GSC分支 Φ_{GSC}以基于POI的GSC向量$\mathbf{g}$ 为输入，并生成自身的逻辑值$\mathbf{z}_{\text{GSC}}$。融合通过对这两个逻辑值进行加权组合来实现。此处不采用单一标量权重，而是使用可学习的、按类别区分的权重向量$\boldsymbol{\lambda} \in \mathbb{R}^C$ 。这种设计为每个类别分配单独的 GSC 权重，同时保持音频分支不变。融合后的逻辑值$\mathbf{z}_{\text{fused}}$ 计算如下：

$$\mathbf{z}_{\text{fused}} = \mathbf{z}_{\text{audio}} + \boldsymbol{\lambda} \odot \mathbf{z}_{\text{GSC}} \tag{1}$$

其中，$\odot$表示逐元素相乘，$\boldsymbol{\lambda}$通过 softplus 激活函数 [41]被约束为非负值，$\boldsymbol{\lambda}$ =softplus($\boldsymbol{\lambda}_{\text{raw}}$)，而 $\boldsymbol{\lambda}_{\text{raw}}$初始化为零。融合在逻辑值（sigmoid 之前）域中进行，其中$\mathbf{z}$表示按类别计算的对数优势分数。因此，公式（1）在最终通过 sigmoid 映射得到概率之前，先结合不同模态的证据。最终类别概率通过对$\mathbf{z}_{\text{fused}}$应用 Sigmoid 函数得到。

GeoFusion-AT 框架在融合后的逻辑值上采用标准的多标签音频标注目标函数。辅助损失和正则化项均为可选项，并非框架定义所必需。

5.2. GeoFusion-AT 框架的实例化

GeoFusion-AT 在三种具有代表性的音频骨干网络上实例化，为 Geo-AT 任务提供基准测试结果。PANNs [20] 是一种基于卷积神经网络的预训练音频骨干网络，AST [9]是一种基于块的 Transformer 骨干网络，可在频谱图块嵌入上应用注意力机制，而 CLAP [21] 是一种通过对比学习预训练的音频-文本骨干网络。所有实例化均遵循第 5.1 节中的定-义：特征级融合（GeoFusion-Early）、表示

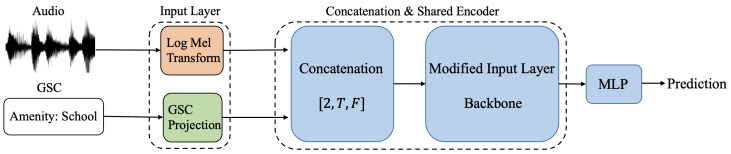


图5：GeoFusion-Early的实例化（特征级融合）。

级融合（GeoFusion-Inter）和决策级融合（GeoFusion-Late）。源代码和模型检查点可在项目主页获取。

*5.2.1. GeoFusion-Early*的实例化

GeoFusion-Early通过构建声学表示张量和广播 GSC 张量来实现特征级融合，如图5所示。

GeoFusion-Early-PANNs。在PANNs上的实例化 [20]遵循第5.1.1节。GSC向量$\mathbf{g} \in \mathbb{R}^{D_{\mathrm{GSC}}}$（含 $D_{\mathrm{GSC}} = 768$）被线性投影为长度F的向量，并沿时间广播，形成广播 GSC 张量$\mathbf{G} \in \mathbb{R}^{1\times T\times F}$。融合前对$\mathbf{A}$应用音频预处理操作。融合后的输入为$\mathbf{X}_{\mathrm{fused}} = \mathrm{Concat}(\mathbf{A}, \mathbf{G}) \in \mathbb{R}^{2\times T\times F}$。第一卷积层经过调整以接受两个输入通道。音频通道的权重由PANNs检查点初始化，GSC通道的权重采用零初始化，以便在初始化时保留预训练音频通路，并让模型在微调期间学会使用$\mathbf{g}$。

GeoFusion-Early-AST。对于 AST，[9], GeoFusion-Early 在词元序列中以特征级融合的方式实现。GSC 向量 $\mathbf{g}$ 不会按通道拼接，而是映射到 AST 嵌入维度，并作为专用的 [GSC] 词元注入。Transformer 输入序列包含标准的 [CLS]词元、[GSC] 词元和音频块词元。位置嵌入表扩展为 $(1, N_{\mathrm{patches}} + 2, D_{\mathrm{emb}})$（其中$D_{\mathrm{emb}} = 768$），新增的 [GSC] 位置初始化为零，而原有位置则保留 AST 检查点中的预训练值。分类使用 [CLS] 词元的输出嵌入。

GeoFusion-Early-CLAP。CLAP 音频编码器 [21] 接受频谱图输入，其实例化时采用与 GeoFusion-Early-PANNs 相同的双通道构造方式。由**g**构造广播 GSC 张量 $\mathbf{G}$，并与 $\mathbf{A}$ 拼接形成 $\mathbf{X}_{\text{fused}}$。音频通道的权重从检查点初始化，而 GSC 通道则初始化为零，以避免在训练初期扰动预训练音频表示。

5.2.2. GeoFusion-Inter 的实例化

GeoFusion-Inter 是一种表征级融合变体，使用第5.1.2节中的对称跨模态注意力模块，将音频嵌入$\mathbf{E}_{\text{audio}}$ 与 GSC 嵌入$\mathbf{E}_{\text{GSC}}$ 相结合，如图 6 所示。

GeoFusion-Inter-PANNs。对于 PANNs，其预训练骨干网络充当特征提取器，用于生成音频嵌入 $\mathbf{E}_{\text{audio}} \in \mathbb{R}^{D_{\text{emb}}}$ （$D_{\text{emb}} = 2048$对于 PANNs）。与此同时，GSC向量 $\mathbf{g} \in \mathbb{R}^{D_{\text{GSC}}}$ 通过两层多层感知机投影为 GSC 嵌入$\mathbf{E}_{\text{GSC}} \in \mathbb{R}^{D_{\text{emb}}}$。使用第5.1.2节中的对称跨模态注意力模块对这些嵌入进行组合，生成$\mathbf{E}_{\text{fused}}$，并将其输入分类头以输出多标签logits。

GeoFusion-Inter-AST。对于 AST，使用 [CLS]输出嵌入作为$\mathbf{E}_{\text{audio}} \in \mathbb{R}^{D_{\text{emb}}}$，其中 $D_{\text{emb}} = 768$。GSC向量 $\mathbf{g} \in \mathbb{R}^{D_{\text{GSC}}}$ 具有 $D_{\text{GSC}} = 768$，并用于构成$\mathbf{E}_{\text{GSC}}$，使其维度保持一致。第5.1.2节中的注意力模块生成用于音频标注的$\mathbf{E}_{\text{fused}}$。

GeoFusion-Inter-CLAP。对于 CLAP，其音频编码器会生成$\mathbf{E}_{\text{audio}} \in \mathbb{R}^{D_{\text{emb}}}$，其中包含 $D_{\text{emb}} = 1024$。与此同时，两层多层感知机将 GSC向量**g**映射为匹配的 GSC 嵌入$\mathbf{E}_{\text{GSC}}$。其中的注意力模块

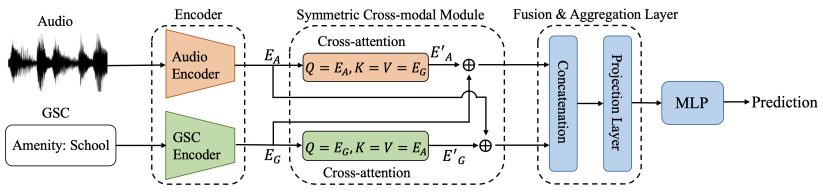


图 6：GeoFusion-Inter 的实例化（表征级融合）。

第5.1.2节将这些嵌入结合起来，生成$\mathbf{E}_{\text{fused}}$，用于标注。

*5.2.3. GeoFusion-Late*的实例化

GeoFusion-Late通过结合音频逻辑值$\mathbf{z}_{\text{audio}}$与GSC逻辑值$\mathbf{z}_{\text{GSC}}$，使用公式1实现决策级融合，如图7所示。

GeoFusion-Late-PANNs。音频分支为PANNs模型，并输出$\mathbf{z}_{\text{audio}}$。GSC分支是一个多层感知机，将$\mathbf{g}$映射到$\mathbf{z}_{\text{GSC}}$。融合逻辑值$\mathbf{z}_{\text{fused}}$通过公式1计算，并使用多标签目标函数进行优化。

GeoFusion-Late-AST。音频分支为AST模型，并输出$\mathbf{z}_{\text{audio}}$。GSC分支与逻辑值融合遵循GeoFusion-Late-PANNs，生成$\mathbf{z}_{\text{fused}}$用于多标签标注。

GeoFusion-Late-CLAP。音频分支使用CLAP 音频编码器生成音频嵌入，再经由分类头输出$\mathbf{z}_{\text{audio}}$。GSC分支与逻辑值融合遵循GeoFusion-Late-PANNs。

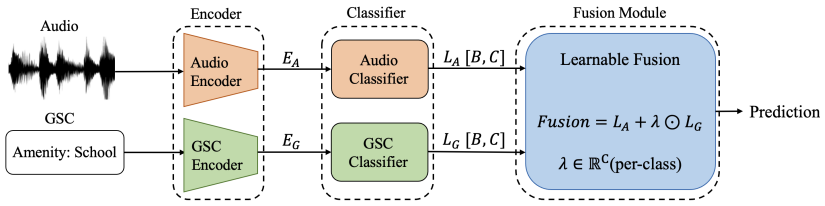


图7：GeoFusion-Late（决策级融合）的实例。

6. 实验与结果分析

6.1. 实验设置与评估指标

Geo-ATBench 被设定为一项 28类多标签音频标注任务；每个音频片段均由声学输入和配对的基于POI的GSC 表征，其内容见第3节和第4节。为进行重复评估，采用不同随机种子进行五次独立运行。每次运行中，数据集按 70% 训练集、15% 验证集和 15% 测试集的比例划分。采用多标签分层方法，以确保各划分中的各标签出现率和共现模式

具有可比性，从而确保测试集中涵盖所有事件类别。划分以片段级为单位进行。GSC 表征并非根据 GPS坐标、街道名称或地址等精确地理标识构建，而是编码从附近兴趣点中提取的高层语义上下文。具体而言，提取原始 OSM标签（例如“设施：学校”和“道路：公交车站”），并将其转换为描述性字符串。随后，使用预训练的 BERT 模型对这些字符串进行编码 [42], ，并对嵌入进行逐元素平均池化，以捕捉当地土地利用特征和区域语义。不同录音位置可能具有相似的 GSC 模式，而同一区域内的录音也可能在兴趣点构成上有所不同。因此，所报告的基准测试结果应理解为评估在片段级划分且包含基于位置的语义上下文时的泛化能力，而非严格地理留出条件下的泛化能力。

本文使用的三种骨干网络（PANNs [20], AST [9], 和 CLAP [21]）均在大规模 AudioSet [26] 上进行预训练，并且在提出时，已报告其在包含 527 个音频事件类别的 AudioSet 上表现出色。在 Geo-ATBench 上构建 Geo-AT 基准时，通过微调使这些骨干网络适配 28 类多标签任务，同时尽量减少对其预训练音频表示的改动。为此，采用较小的学习率和早停策略。模型在 NVIDIA GeForce RTX 4090 GPU 上训练，使用 AdamW优化器，以 1e-5 的学习率进行微调，最多训练 100 个周期。为防止过拟合，如果验证集F1分数连续 15 个周期没有提升，则停止训练。训练目标为二元交叉熵（BCE）损失 [20]。音频输入为 10 秒的音频片段，并重采样以符合各骨干网络的要求。所有模型均使用预训练权重初始化，并纳入仅音频基线进行比较。

模型性能通过平均精度均值（mAP） [20],进行评估

ROC曲线下面积（ROC AUC）和 F1 分数进行评估，并报告 5 次独立运行的平均值 ± 标准差。除非另有说明，所有指标均采用微平均。除针对 28 个事件类别的多标签音频标注外，还报告了 3 类粗粒度音频标注，作为补充分析。有关所有源代码、模型和数据集，请参阅项目主页。

6.2. 结果与分析

本节从三个互补的角度评估 Geo-ATBench。首先，在不同的 POI 提取范围下评估仅凭 GSC 进行音频标注的可行性，并将其作为仅 GSC 基线。其次，报告三种基于 AudioSet 预训练的强大音频骨干网络的仅音频零样本基线，以刻画这些骨干网络在 Geo-ATBench 上微调之前的表现。第三，报告在相同数据划分下，Geo-ATBench 上仅音频和 GeoFusion-AT 变体的微调 Geo-AT 结果，从而能够对特征级、表征级和决策级融合进行受控比较。通过逐标签性能变化和错误模式，识别哪些标签及混淆最能从 GSC 中获益，尤其关注声学易混淆标签。

6.2.1. 仅 *GSC* 基线与 *GSC*范围敏感性

实际中，不同声音事件的可感知范围各异，其与附近地点语义的相关程度也不尽相同。因此，POI 提取范围会影响构建 GSC向量 g 时可用的 POI派生上下文的数量与组成。如图 8 所示，为了评估 Geo-ATBench 上的纯地理基线，我们在多个 POI提取范围下评估了仅 GSC 基线。

对于每个由距离阈值定义的 POI提取范围（以正方形边长实现），都以音频片段的 GPS坐标为中心构建一个正方形。尽

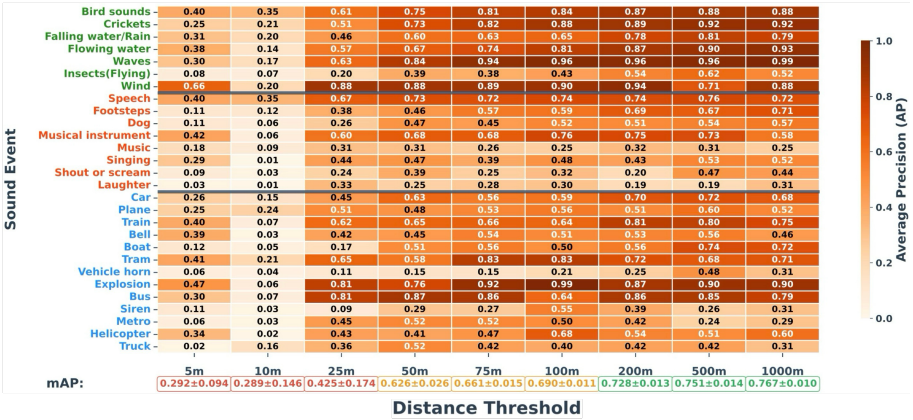


图 8：不同 POI 提取范围下仅使用 GSC 的多标签标注平均精度（AP）。基于 POI 的 GSC 由使用相同的 11 个 OSM 要素键（例如土地利用、便利设施）检索到的 OSM 实体构建，并编码为定长GSC向量g。平均精度均值（mAP）是在测试集上经过 5 次独立运行计算得出的。

管圆形邻域或许能更好地近似各向同性的声音传播，但为便于在基于 OSM 的地理信息系统中高效执行边界框查询，我们采用正方形区域。使用与第4节相同的 11 个 OSM要素键 [36]检索正方形内的 OSM实体。随后，使用预训练的 BERT 模型 [42]将得到的兴趣点构成编码为固定长度的 768 维 GSC向量 g，并在所有范围内评估同一个仅GSC分类器。仅使用地理信息的基线使用 BERT-base 生成 g，再通过由 1024、512 和 28 个单元组成的 3 层多层感知机，对 Geo-ATBench 执行 28 类多标签标注。训练期间，BERT 分词器和 BERT 编码器 [42] 保持冻结，仅微调 3 层多层感知机分类器。源代码、提取的 GSC向量以及实现细节均可在项目主页上获取。

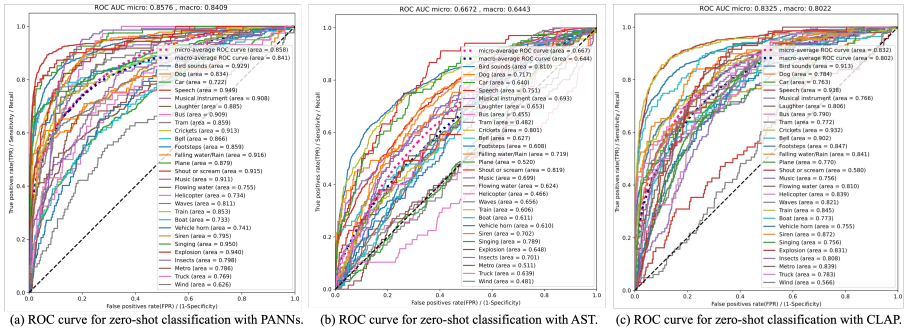
在 Geo-ATBench 上，仅GSC的结果随着距离阈值增大而提升，1000 米范围取得了最高性能。一种可能的解释是，某些地区的 OpenStreetMap（OSM）覆盖较为稀疏 [43], ，因此较小的方形区域可能只能检索到较少的实体来构建 GSC向量 g。不同设备和使用条件下的 GPS 精度也可能有所不同 [44],

这可能会使查询区域发生偏移，并影响兴趣点检索。其他因素也可能有所影响，因为不同声音事件的语义与附近地点的关联范围，以及它们与 POI派生上下文的契合程度各不相同。例如，鸟类或蟋蟀等移动声源的声音可以在自然区域内广泛传播，并可能与录音点附近以外的林地或水域兴趣点相关联。人类语音也可能反映附近的景点或行人流：人们可能向某个地点靠近或远离，并在较大范围内产生语音声响。相比之下，与固定声源相关的事件（例如海岸线上的破碎波）通常局限于更局部的地点语义范围；因此，在某些情况下，较短的提取范围就足够了。总之，本节展示了 Geo-ATBench 在不同兴趣点提取范围下的仅地理信息性能，为所提出的 Geo-ATBench数据集提供了详细的仅地理信息对比参考。

6.2.2. 仅音频零样本基线

报告仅音频零样本标签推断结果，以刻画三个基于 AudioSet 预训练的模型在 Geo-ATBench 上微调之前的表现。通过直接推理对 AudioSet 预训练骨干网络 PANNs [20],AST [9],和 CLAP [21] 进行评估，为后文报告的微调后仅音频结果和 GeoFusion-AT 结果提供参考。

无法直接根据原始 AudioSet 预训练模型的输出，对 Geo-ATBench 标签进行零样本基准测试，因为这些骨干网络是在包含 527 个类别事件标签的 AudioSet 上训练的，其原生输出与 28 个 Geo-ATBench 事件标签并不匹配。为定义一个可比较的 28 标签零样本基准测试，首先需要针对每个 Geo-ATBench 音频片段，在 527 个 AudioSet 标签上生成类别预测，然后使用预训练的 Word2Vec 模型（“word2vec-google-news-300”） [45], 将这 527 个输出映射到 28 个 Geo-ATBench 标签。该模型提供在 Google News语料库上训练的 300 维词嵌入，



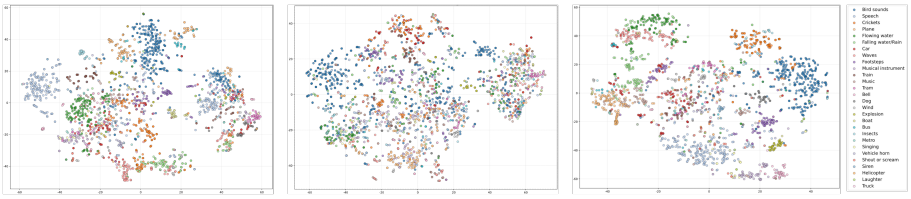
(a) ROC curve for zero-shot classification with PANNs. (b) ROC curve for zero-shot classification with AST. (c) ROC curve for zero-shot classification with CLAP.

图 9：在 Geo-ATBench 标签上进行零样本纯音频标签推理的 ROC曲线。报告了 AudioSet 预训练的 PANNs、AST 和 CLAP 在经过 AudioSet到Geo-ATBench的标签映射后的微平均和宏平均 ROC AUC。总体性能：PANNs（微平均 AUC 0.8576，宏平均 AUC 0.8409）、AST（0.6672，0.6443）和 CLAP（0.8325，0.8022）。

AudioSet 到 Geo-ATBench 的

标签映射和代码已发布在项目主页上，供复现使用。

图 9 展示了三种经过 AudioSet预训练配置的骨干网络在 Geo-ATBench 上的仅音频零样本标注性能。PANNs 取得了最强的表现，其次是 CLAP，而在采用相同的 AudioSet到Geo-ATBench的标签映射时，AST 的表现最差。这一排序可能由多种因素造成。首先，PANNs 经过训练，能够根据对数梅尔输入 [20], 生成较强的音频片段级标签预测；在标签空间映射下，这些预测可以更直接地迁移到 Geo-ATBench。其次，CLAP 学习了对齐的音频-文本表示 [21], ，能够保留语义区分度；将 AudioSet 标签映射到 Geo-ATBench 标签后，这种区分度仍然具有实用价值。第三，AST 依赖频谱图块标记化和位置嵌入 [9], ，其 AudioSet预训练配置在不进行特定任务适配、直接推理的情况下，可能无法有效迁移到 Geo-ATBench 的数据分布。类似地，在图 10 中，对零样本推理下音频嵌入的可视化也显示出相同趋势。PANNs 和 CLAP 的嵌入在 28 个 Geo-ATBench类别之间形成了更易分离的簇，而 AST 的嵌入则表现出更明显的重叠，并集中在空间中较小的区域内。受篇幅限制，图 9 和图 10 的高分辨率版本可在



(a) t-SNE visualization of PANNs audio embeddings. (b) t-SNE visualization of AST audio embeddings. (c) t-SNE visualization of CLAP audio embeddings.

图 10：PANNs、AST 和 CLAP 在 28 个 Geo-ATBench 事件类别上进行零样本推理得到的音频嵌入的 t-SNE 可视化；聚类显示出有效的分离，而重叠区域则突出显示了声学上相似的类别。

项目主页上查看。

6.2.3. Geo-AT 在 Geo-ATBench 上的微调结果

表 2 展示了第 5 节所述三种融合策略下音频骨干网络的结果。同时报告了仅音频和仅 GSC 基线，并在相同的数据划分下，对 GeoFusion-AT 的不同变体在包含 28 个类别的 Geo-AT 任务上进行了比较。所有微调模型均基于 28 个 Geo-ATBench 标签进行训练，未使用零样本基线中采用的 AudioSet 标签映射。

在早期融合、表示层融合和后期融合中，引入 GSC 均能提升三种骨干网络在 28类 Geo-AT 任务上的性能。在五次运行中，Welch双样本t检验表明，与相应的仅音频基线相比，AST 采用早期融合（$p < 0.05$）、PANNs 采用后期融合（$p < 0.001$）以及 CLAP 采用中间融合（$p < 0.01$）均取得了显著提升。GeoFusion-Early-AST 在细粒度

策略	**细粒度（28 个类别）**			**粗粒度（3 个类别）**		
	PANNs	AST	CLAP	PANNs	AST	CLAP
仅音频	0.770±0.006	0.820±0.015	0.824±0.008	0.961±0.006	0.904±0.012	0.966±0.005
仅GSC	0.767±0.010			0.867±0.009		
GeoFusion-Early	0.812±0.010	0.846±0.010	0.826±0.010	0.954±0.004	0.914±0.009	0.950±0.008
增益 *(Δ)*	+0.042	+0.026	+0.002	-0.007	+0.010	-0.016
GeoFusion-Inter	0.824±0.010	0.829±0.003	0.842±0.006	0.964±0.008	0.912±0.011	0.968±0.005
增益 *(Δ)*	+0.054	+0.009	+0.018	+0.003	+0.008	+0.002
GeoFusion-Late	0.833±0.007	0.843±0.010	0.831±0.007	0.949±0.004	0.939±0.008	0.966±0.004
增益 *(Δ)*	+0.063	+0.023	+0.007	-0.012	+0.035	0.000

表 2：不同模型在 Geo-ATBench数据集上的平均精度均值（mAP）。标记为 增益 *(Δ)* 的行表示相较于纯音频基线的性能差异。所有指标均为 5 次独立实验的平均值。

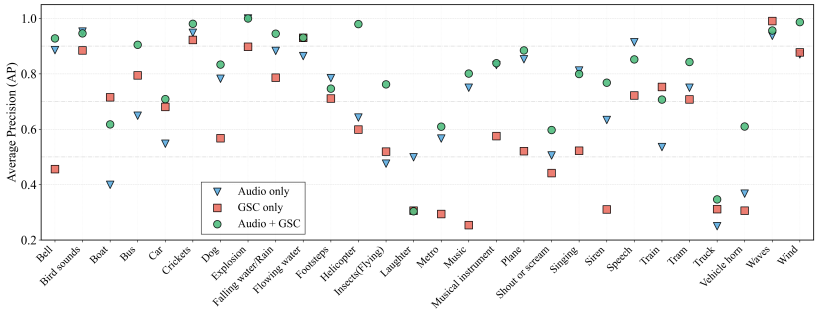


图 11：纯音频 AST、仅使用 GSC 的基线模型和 GeoFusion-Early-AST 在 28 个类别上的逐类平均精度。集成地理空间信息可提升多个类别的性能。

28类多标签标注任务中取得了最佳平均精度均值，而 GeoFusion-Early-AST 与 GeoFusion-Inter-CLAP（$p > 0.5$）之间未观察到显著差异，表明二者性能相当。微调后，AST 在 28类 Geo-AT 任务上取得了最强的总体性能，其次是 CLAP，最后是 PANNs。这一排序不同于图 9 中的零样本基线排名。其主要区别在于，零样本基线在 527 类 AudioSet 标签空间中进行预测，然后将输出映射到 28 个 Geo-ATBench 标签；而经过微调的模型则直接基于 Geo-ATBench 标签进行训练。映射步骤可能引入因骨干网络而异的标签聚合效应和校准效应，而跨数据集领域偏移又进一步限制了直接推理下的直接迁移。微调通过直接针对 Geo-ATBench 目标进行优化，消除了标签空间不匹配问题，从而取得了更好的结果。图 11 进一步展示了 GeoFusion-Early-AST 在 Geo-ATBench 各事件类别上的平均精度。

除 28 类细粒度标注任务外，表 2 还展示了一项补充的 3 类粗粒度标注任务，该任务将第 4.2 节所述的 28 种事件归为自然声音、人类声音和物体声音。在这项粗粒度任务中，GeoFusion-Inter-CLAP 的平均精度均值最高。对于所有三种骨干网络，表征级融合都提升了粗粒度任务的性能，这表明在 GeoFusion-Inter 框架中，利用对称跨模态注意力机制结合音频和 GSC 的高层表征

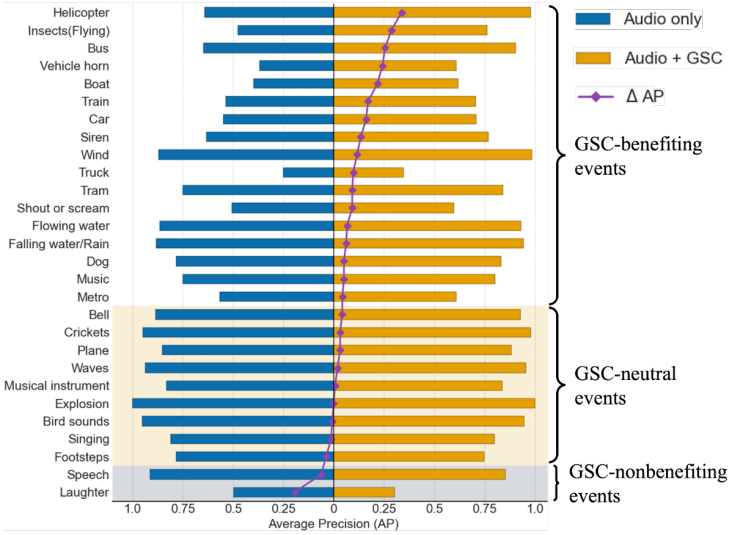


图12：GeoFusion-Early-AST 示例在 28类 Geo-AT 任务上的各类别平均精度变化（融合减去仅音频）。

该方法在这一语义粒度下是有效的。

为进一步探究在 Geo-AT 任务中，哪些事件标签最能受益于引入地理空间语义上下文（GSC），图 12 以 GeoFusion-Early-AST 为例，展示了仅音频变体和音频-GSC 融合变体在各标签上的平均精度（AP）。图 12 还展示了各标签的变化 $\Delta AP = AP_{\text{audio+GSC}} - AP_{\text{audio}}$，以紫色曲线表示。与仅音频基准相比，引入 GSC 后，28 个事件类别中有 17 个的 AP 提高了 5% 以上。本文将这 17 个类别归为受益于 GSC 的事件。其中，直升机 的增益最大，绝对变化为 $\Delta AP = 0.3378$，相较于仅音频 AP，增幅约为 $+52.62\%$。这与直升机声音通常出现在有限的地点、且往往与特定的位置语义相关这一事实相符，因此基于兴趣点的 GSC 有助于消歧。在 28 个事件类别中，有 9 个类别的 ΔAP 保持在 $\pm 5\%$ 以内；这些类别被归为 GSC 中性事件，例如铃、歌唱和脚步声，它们都是常见的日常声音，并且属于

与特定地点语义的关联较弱。值得注意的是，爆炸在该数据集中几乎没有变化；这一模式与从 Freesound.org [35]检索到的爆炸样本相符，这些样本主要涉及烟花等日常活动。最后，有两个类别——语音 和笑声——降幅低于 -5%，因此被归为GSC无益事件。这可能是因为语音和笑声广泛分布于各种地点，将它们与基于POI的GSC关联起来并无助于识别。总体而言，图12表明，GSC有助于识别大多数声音事件类别，对部分常见声音事件类别的影响有限，而对于某些不对应特定地点、广泛存在的人类发声事件，可能没有帮助。

7. Geo-ATBench数据集的人工评估

为评估基于Geo-ATBench数据集训练的模型与人类听觉判断的一致程度，我们开展了一项众包人类听音研究。本研究考察模型预测与人类多标签事件判断之间的对应关系。根据收集到的标注，(1) 使用描述性一致性度量和机会校正信度统计来概述标注一致性；(2) 在片段级将模型预测与聚合人类共识标签进行比较，以评估模型与人类的一致性。

7.1. 参与者与实验设计

十名中国参与者（3名女性、7名男性；平均年龄 $\mathrm{M} = 22.4$ 岁，标准差 $\mathrm{SD} = 0.70$ 岁）参加了评估实验。参与者的语言背景相似，以确保他们能一致地理解标注界面和任务说明。本研究遵循利物浦约翰摩尔斯大学的伦理准则，并已获得所有参与者的知情同意。

为评估 Geo-ATBench 标签及下游模型预测的感知有效性，我们开展了一项被试内人工标注实验。参与者聆听 579 个 Geo-ATBench 音频片段，并在清晰听到相应事件时选择“存在”，以标记该事件是否存在。单个音频片段中可以标记多个事件为存在，这与 Geo-AT 中的多标签标注形式一致。每位参与者都会收到全部音频片段，以及一份与 28 个 Geo-ATBench 事件标签相对应的事件类别清单。参与者可以反复播放音频片段，直至做出有把握的判断。研究要求参与者依据听觉感知进行判断，与第 4.2 节所述的 Geo-ATBench 标注流程保持一致。标注任务分为若干短时段，并以随机顺序呈现，以减少疲劳和顺序效应。每位参与者须在 14 天内完成所有音频片段的标注。

7.2 人类聆听研究结果：可靠性与模型和人类的一致性 *7.2.1* 人类多标签事件标注的评定者间信度

针对每位参与者，计算其与聚合人类共识标签之间的描述性一致率。所有参与者的平均一致率为 0.97，表明参与者在不同音频片段和声音事件类别上的判断相似。标注矩阵较为稀疏，只有约 4.5% 的片段–事件位置被标记为 1（存在）。这种类别不平衡会抬高原始百分比一致率，因为大多数标注都属于同一类别（不存在）。为获得经过机会校正的可靠性估计，计算名义数据的克里彭多夫阿尔法系数。将每个片段–事件对视为一个条目，十位参与者共计产生 16,212 个条目。所得信度系数为 $\alpha_{\text{nominal}}(N = 16,\ 212,\ \ R = 10) = 0.486$，表明十位参与者之间的一致性中等，也说明不同个体在多标签复音声音事件的听觉感知和解释方面存在差异

。鉴于一致性中等，我们采用多数投票，为每个事件生成片段级共识标签，作为模型比较的参考。数据收集完成后，为每位参与者生成二元事件矩阵，并按片段–事件对进行汇总：当十位参与者中至少有五位标记为“存在”时，赋值为 1；否则赋值为 0。

总体而言，标注显示出较高的原始一致性，但经机会水平校正后的可靠性仅为中等；在稀疏的二元多标签标注中，这一结果是预期之内的。多数投票得出的共识标签被用作音频片段级别的参考，以衡量模型与人类的一致性；对于低流行率事件，则需谨慎解读。下一小节将比较模型预测与聚合人类共识标签，以量化模型在 Geo-ATBench 上与人类的一致性。

7.2.2. 两种标签参考下模型与人类的一致性

为评估模型评价对标签参考选择的敏感程度，我们针对同一测试集中的 579 个音频片段，使用两组标签评估模型预测：（i）Geo-ATBench 标签；（ii）来自 10 位参与者的聚合人类共识标签。共识阈值设为 0.5，即当至少 5 位标注者将某事件标记为存在时，该事件便被视为存在。本比较旨在评估 Geo-ATBench 标签参考是否与独立的人类判断一致。

报告了仅音频 CLAP 基线和 Geo-AT 变体 GeoFusion-Inter-CLAP 的结果；该变体在包含 28 个类别的细粒度 Geo-AT 任务和包含 3 个类别的粗粒度 Geo-AT 任务中表现出色。针对两种标签参考下的 28 个逐类 F1 分数进行了配对 Wilcoxon 符号秩检验。结果显示，对于仅音频 CLAP，其在 Geo-ATBench 标签与聚合人类共识标签下的表现不存在统计显著差异；GeoFusion-Inter-CLAP 也得出了相同结论。具体而言，配对 Wilcoxon 符号秩

事件	F1（标签）	F1（人工）	时长	事件	F1（标签）	F1（人工）	时长
鸟类声音	0.861	0.856	8191	落水声	0.872	0.786	2922
语音	0.827	0.869	5133	流水	0.774	0.667	2774
飞机	0.779	0.514	3092	波浪	0.756	0.677	2754
蟋蟀	0.883	0.836	3091	脚步声	0.629	0.607	2225
Car	0.491	0.404	3068	乐器	0.651	0.667	1593

表 3：Geo-ATBench 中按总时长降序排列的前 10 个事件类别（总计 $=34{,}843$ 秒，≈ 9.68 小时）。F1（标签）和 F1（人工）分别表示 GeoFusion-Inter-CLAP 预测结果与 Geo-ATBench 标签及聚合人类共识标签进行评估所得的 F1 分数。时长表示事件总时长（秒）。乐器指乐器，落水声对应落水声/雨。

检验表明，仅音频 CLAP 的 F1 分数在 Geo-ATBench 标签（$W = 181$，$p > 0.05$）与 10 位参与者的人类共识标签（$F1 = 0.628$）之间没有显著差异。为简洁起见，表 3 报告了总时长最长的 10 个事件类别的逐类 F1 分数，而统计检验使用了全部 28 个类别。GeoFusion-Inter-CLAP 也呈现出类似的模式：在 Geo-ATBench 标签（$F1 = 0.649$）和 10 位参与者的人类共识标签（$F1 = 0.592$；$W = 187$，$p > 0.05$）下，F1 分数均保持稳定。总体而言，在包含 579 个音频片段的标注测试集上，模型评估结果在 Geo-ATBench 标签和聚合人类共识标签下保持一致；配对 Wilcoxon 符号秩检验也未显示两种标签参考之间存在统计显著差异。这一结果补充了评定者间信度分析，并支持将 Geo-ATBench 作为与人类判断一致的片段级评估基准。

8. 结论

环境声音事件并非孤立存在：它们是在特定地理环境中产生并被感知的物理现象。然而，计算听觉场景分析（CASA）通常将多标签音频标注视为仅音频推断问题。在复音重叠的情况下，音频信息可能不足以区分具有相似声学模式的标签，而消除歧义的线索可能存在于波形之外。

为弥补这一不足，我们提出地理空间音频标注（Geo-AT）任务，将多标签音频标注定义为基于配对音频与地理空间语义上下文（GSC）的任务。Geo-AT 着重利用源自兴趣点、与位置相关的语义作为补充先验，这些先验并未编码在音频信号中。这一任务层面的定义为将空间语义融入机器听觉提供了原则性基础，使计算听觉场景分析（CASA）的范围超越独立的信号分析。

Geo-ATBench 的发布旨在支持可复现的 Geo-AT 评估。Geo-ATBench 包含 3,854 个音频片段（10.71 小时），涵盖真实世界复音音频的 28 个事件类别，且每个音频片段都配有基于 OSM 标注、覆盖 11 个语义上下文类别构建的基于POI的GSC表示。通过显式编码录音环境的语义特征，Geo-ATBench 填补了该领域一项重要的资源空白。它使研究者能够开展受控研究，考察空间上下文如何与声学表示相互作用，并为评估地理空间接地的声音分类提供共享基准。本文提出 GeoFusion-AT，以在 Geo-ATBench 上报告参考结果，采用特征级、表示级和决策级融合，并使用三个经 AudioSet 预训练的骨干网络：PANNs、AST 和 CLAP。在不同骨干网络与融合点上，引入 GSC 在大多数配置中均与 28类 Geo-AT 任务上平均精度均值的提升相关。

一项由 10 位参与者参与的众包听音研究进一步证实，Geo-ATBench 可作为与人类判断一致的参考基准，用于包含 579 个音频片段的标注测试集。Geo-AT 任务、Geo-ATBench数据集和 GeoFusion-AT 框架共同为研究位置相关语义如何在 CASA 中补充声学表示提供了具体基础。

参考文献

[1] T. Virtanen、M. D. Plumbley、D. Ellis，《声音场景与事件的计算分析》，第 9 卷，施普林格，2018。

[2] A. Mesaros, T. Heittola, E. Benetos, P. Foster, 等，《声学场景与事件的检测和分类：DCASE 2016 挑战赛成果》，IEEE《音频、语音与语言处理汇刊》26 (2) (2017年) 379–393。

[3] K. Lopatka, 等，《存在背景噪声时声学事件的检测、分类与定位：用于危险情境声学监控》，《多媒体工具与应用》75(17) (2016) 10407–10439。

[4] D. Zhang, Z. Zhong, 等，《智慧城市环境声音自动分类系统》，《传感器》23 (15) (2023年) 6823。

[5] E. Wold, T. Blum, D. Keislar, J. Wheaton，《音频的基于内容的分类、搜索与检索》，IEEE《多媒体》3 (3) (1996) 27–36。

[6] M. Vacher、F. Portet 等，《用于智能家居非侵入式事件检测的音频内容分析》，Eng. Appl. Artif. Intell. 89 (2020年) 103226。

[7] A. Mesaros、T. Heittola、T. Virtanen，《声音事件检测：教程》，IEEE Signal Process. Mag. 38 (5) (2021年) 67–83。

[8] H. Purwins、B. Li、T. Virtanen 等，《用于音频信号处理的深度学习》，J. Acoust. Soc. Am. 146 (6) (2019年) 4392–4395。

[9] Y. Gong、Y. A. Chung、J. Glass，《AST：音频频谱图Transformer》，载于：Proc. Interspeech，2021年，第571–575页。

[10] J. Salamon, J. P. Bello，面向环境声音分类的深度卷积神经网络与数据增强，《IEEE 信号处理快报》24 (3) (2017年) 279–283。

[11] V. Morfi, D. Stowell，低资源数据集上的音频事件检测、标注与深度学习，《应用科学》8 (8) (2018) 1397。

[12] T. Heittola、A. Mesaros 等，《依赖上下文的声音事件检测》，EURASIP 音频、语音与音乐处理期刊 2013年 (1) (2013年) 1–13。

[13] P. Primus、G. Widmer，《融合音频和元数据嵌入可改进基于语言的音频检索》，载于：EUSIPCO 会议论文集，2024年，第 321–325 页。

[14] A. S. Bregman，《听觉场景分析：声音的感知组织》，麻省理工学院出版社，1994。

[15] I. Diez、I. Saratxaga、U. Salegi 等，《NoisenseDB：用于开发噪声监测应用神经分类系统的城市声音事件数据库》，《应用科学》13 (16) (2023年) 9358。

[16] J. Yuan 等，《利用人类移动和兴趣点发现城市中不同功能区域》，载于：ACM SIGKDD 会议论文集，2012年，第 186–194 页。

[17] D. K. Dalei, S. Shrivastava, N. Panigrahi，从地理地点到语义空间：地理空间嵌入分析，ISPRS 摄影测量、遥感与空间信息科学年刊 10 (2025) 131–135。

[18] M. Cartwright 等，SONYC-UST-V2：具有时空上下文的城市声音标注数据集，载于：DCASE 会议论文集，2020年，第 16–20 页。

[19] A. Kumar, B. Raj，使用弱标注数据进行音频事件检测，载于：ACM 多媒体会议论文集，2016 年，第 1038–1047 页。

[20] Q. Kong、Y. Cao、T. Iqbal 等，PANNs：用于音频模式识别的大规模预训练音频神经网络，《IEEE/ACM 音频、语音与语言处理汇刊》28（2020年），2880–2894页。

[21] B. Elizalde、S. Deshmukh、M. Al 等，CLAP：通过自然语言监督学习音频概念，载于：ICASSP 会议论文集，2023年，第 1–5 页。

[22]E. Fonseca 等，FSD50K：一个人工标注的声音事件开放数据集，《IEEE 音频、语音与语言处理汇刊》30（2021年）829–852。

[23]K. J. Piczak，ESC：环境声音分类数据集，载于：《ACM 多媒体》会议论文集，2015年，第1015–1018页。

[24] M. Sandler、A. Howard、M. Zhu 等，MobilenetV2：倒置残差与线性瓶颈，载于：CVPR 会议论文集，2018，第4510–4520页。

[25] K. Chen、X. Du、B. Zhu、Z. Ma 等，HTS-AT：用于声音分类与检测的分层令牌语义音频 Transformer，载于：ICASSP 会议论文集，2022年，第646–650页。

[26] J. F. Gemmeke、D. P. Ellis 等，Audio Set：用于音频事件的本体与人工标注数据集，载于：ICASSP 会议论文集，2017年，第776–780页。

[27] H. Chen, W. Xie, A. Vedaldi, A. Zisserman，VGGSound：一个大规模音视频数据集，载于：ICASSP 论文集，2020年，第 721–725 页。

[28] C. D. Kim, B. Kim, H. Lee, G. Kim，AudioCaps：为真实场景中的音频生成描述，载于：NAACL-HLT 论文集，2019年，第 119–132 页。

[29] Y. Zhou, Z. Wang, C. Fang 等，视觉转声音：为真实场景中的视频生成自然声音，载于：CVPR 论文集，2018，第 3550–3558 页。

[30] P. Yang, X. Wang, X. Duan 等，AVQA：一个针对视频音视频问答的数据集，载于：ACM 多媒体论文集，2022年，第 3480–3491 页。

[31] M. B. Shaikh, D. Chai 等，《用于音频-图像融合及视频动作识别的多模态融合》，Neural Comput. Appl. 36 (10) (2024年) 5499–5513。

[32] C. A. Quinn, P. Burns, G. Gill 等，《卷积神经网络声景分类揭示生态声学数据中的时间和地理模式》，Ecol.Indic. 138(2022年)108831。

[33] M. Boes、K. Filipan 等，《用于公园声景质量评估的机器听觉》，《声学学报》104（1）（2018）121–130。

[34] E. K. Wang、H. Liu、G. Wang、Y. Ye 等，《面向自适应助听器的情境识别》，载于：INDIN 会议论文集，2015年，第 1102–1107 页。

[35] E. Fonseca、J. Pons、X. Favory 等，《Freesound 数据集：创建开放音频数据集的平台》，载于：ISMIR 会议论文集，2017年，第 486–493 页。

[36] OpenStreetMap 贡献者，OpenStreetMap，https://www.openstreet map.org，访问日期：2024-02-10（2024年）。

[37] H. Schütze、C. D. Manning、P. Raghavan，《信息检索导论》，第 39 卷，剑桥大学出版社，剑桥，2008年。

[38] C. G. M. Snoek、M. Worring、A. W. M. Smeulders，《语义视频分析中的早期融合与后期融合》，载于：ACM 多媒体 2005 年会议论文集，第 399–402 页。

[39] J. Ngiam、A. Khosla、M. Kim、J. Nam、H. Lee、A. Y. Ng等，多模态深度学习，载于：Proc. ICML，第11卷，2011年，第689–696页。[40] H. Akbari、L. Yuan、R. Qian等，VATT：面向原始视频、音频和文本的多模态自监督学习Transformer，载于：Proc. NeurIPS，第34卷，2021年，第24206–24221页。[41] X. Glorot、A. Bordes、Y. Bengio，深度稀疏整流神经网络，载于：Proc. AISTATS，2011年，第315–323页。[42] J. Devlin、M. W. Chang、K. Lee等，BERT：面向语言理解的深度双向Transformer预训练，载于：Proc. NAACL-HLT，2019年，第4171–4186页。

[43] B. Herfort、S. Lautenbach等，对OpenStreetMap中全球城市建筑数据完整性与不平等性的时空分析，《自然·通讯》14 (1)（2023年）3985。

[44] P. A. Zandbergen，iPhone 位置的准确性：辅助全球定位系统、WiFi 和蜂窝定位的比较，《GIS 汇刊》13（2009）5–25。

[45] T. Mikolov、K. Chen、G. Corrado、J. Dean，向量空间中词表示的高效估计，载于：国际学习表征会议 ICLR 论文集，2013 年。