声呐时刻：音频语言模型的音频地理定位基准测试

Ruixing Zhang[1], Zihan Liu[1,2],LeileiSun[1,3],Tongyu Zhu[1,3],WeifengLv[1][1]复杂关键软件环境国家重点实验室，北京航空航天大学[2]上海人工智能实验室，[3]数据科学与智能计算重点实验室，北京航空航天大学国际创新研究院

arXiv:2601.03227v1 [cs.SD] 6 Jan 2026

摘要

地理定位旨在推断给定信号的地理来源。在计算机视觉领域，地理定位一直是组合推理的一项高难度基准测试，并与公共安全密切相关。相比之下，音频地理定位的发展受到高质量音频-位置对缺乏的制约。为弥补这一空白，我们推出了 AGL1K，这是首个面向音频语言模型（音频语言模型）的音频地理定位基准测试，覆盖72个国家和地区。为了从众包平台中可靠地提取可本地化样本，我们提出了音频可定位性指标，用于量化每段录音的信息量，最终获得了1,444个精选音频片段。对16个音频语言模型的评估表明，音频语言模型已经具备音频地理定位能力。我们发现，闭源模型的表现显著优于开源模型，且语言线索往往主导预测过程，成为预测的支撑。我们进一步分析了音频语言模型的推理轨迹、区域偏差、错误原因，以及可定位性指标的可解释性。总体而言，AGL1K建立了一个音频地理定位基准测试，有望推动音频语言模型获得更强的地理空间推理能力。

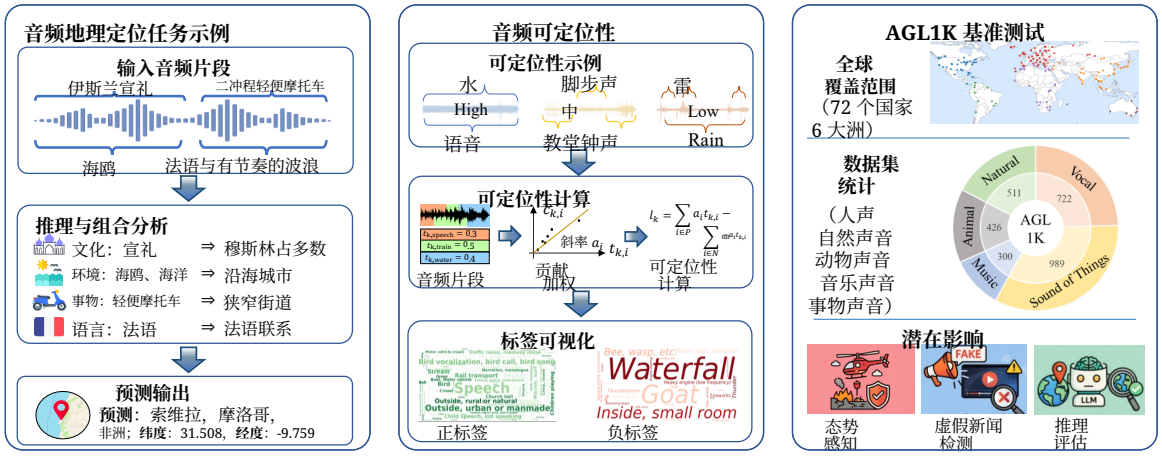
代码： https://github.com/Rising0321/AGL1K
空间： https://huggingface.co/spaces/RisingZhang/AudioGeoLoc

1 引言

地理定位旨在推断信号的地理来源。与标准分类相比，它是一种颇具吸引力的替代方法，因为它需要综合多种线索进行推理。将观测结果映射到单一 GPS 坐标，不仅需要感知推断，还需要关于地理和文化的广泛世界知识。在计算机视觉领域，这一问题通常作为图像地理定位进行研究（Weyand 等人，2016；Regmi 和 Shah，2019；Li 等人，2024年）。至于音频，它也可以提供类似的证据，包括混响模式、交通密度和海岸波浪动力学。除了学术研究价值，音频地理定位还具有明确的社会价值。例如，协助调查极端主义行为者传播的音频，可以帮助事实核查员核实病毒式传播内容所声称的位置。这些应用使音频地理定位与公共安全息息相关。然而，尽管具有这些潜力，**目前仍缺乏针对音频地理定位的系统性基准测试**。

音频地理定位基准测试的缺失主要源于两个因素。首先，**目前没有公开可用且带有位置标注的音频数据集**。与图像地理定位不同，后者的发展得益于社交媒体平台提供的大规模地理标记数据，而音频领域不存在类似的资源。其次，**该领域缺乏对音频位置可定位性的量化概念，因而无法筛选出具有地理信息的录音**。即使众包平台能够提供大量音频-位置对，缺少这样的衡量标准，也仍然难以识别出包含有意义地理信号的样本。

为系统评估音频语言模型（音频语言模型）的音频地理定位能力，我们推出了 AGL1K，这是首个面向音频语言模型的音频地理定位基准测试。AGL1K 基于众包平台 Aporee 平台整理而成，并通过我们提出的音频可定位性指标进行筛选。该指标在推理过程中汇总来自正向和负向声音类别的证据，以估计每段录音所包含的地理信息量。最终形成的基准测试覆盖六大洲的 72 个国家，涵盖多种声学场景，包括自然声景、动物发声、音乐、人造声音和口语对话。**这种多样性使 AGL1K 成为评估现代音频语言模型组合推理能力的理想测试平台，并具有潜在**



Figu图1：AGL1K概览，展示音频地理定位任务、可定位性和组合 n.

对公共安全和虚假信息检测的下游意义。

我们对 16 个音频语言模型进行了全面评估，结果表明，当前的音频语言模型已开始展现出有意义的音频地理定位能力，但 AGL1K 仍颇具挑战性。结果揭示了闭源模型与开源模型之间明显的能力层级差异。领先的音频语言模型（例如 Gemini 3 Pro）展现出强大的知识储备和推理能力，其失败越来越多地并非源于信息缺失，而是过度依赖单一线索。相比之下，开源模型在细粒度感知方面存在更根本的局限，因而无法可靠地提取具有地理信息量的信号。

通过详细的错误分析，我们总结出三项发现，为未来音频模型的改进指明方向。**(1)增强细粒度感知**：许多开源音频语言模型会误判语言及其他细微的声学线索，限制了感知敏感度和下游知识检索。**(2) 缓解区域偏差：** 在不同模型中，系统性的预测失衡现象依然存在，例如对某些大洲或地区的预测偏多。**(3) 加强组合推理**：由于音频地理定位需要整合多个微弱线索，模型必须避免将任何单一线索视为决定性证据。

总之，本研究作出以下贡献：

- **首个面向音频语言模型的音频地理定位基准测试。**我们推出 AGL1K，这是一项基准测试，包含来自 72 个国家的 1,444 个用户上传的音频片段，涵盖多样的声学场景，包括自然声景、动物发声、音乐、人造声音和口语对话。

- **音频可定位性的原则性定义。** 我们提出音频可定位性这一概念，用于量化音频的地理信息量。该指标在推理过程中汇总正向和负向声音标签所提供的信息量，从而能够筛选出可定位的录音。

- **对最先进音频语言模型进行全面评估。**我们对16个音频语言模型进行了基准测试，并在一定程度上发现了音频地理定位的兴起。我们进一步分析了代表性模型的推理轨迹、区域偏差和错误原因，为未来改进音频语言模型提供了见解。

2 相关工作

2.1 地理定位

现有地理定位研究大多聚焦于基于图像的地理定位，大致可分为三种范式。**基于分类的**方法将地球表面划分为预定义区域，但要实现细粒度精度，就需要大量类别，这使得模型训练和扩展面临挑战（Weyand 等人，2016；Clark 等人，2023年；Müller-Budack 等人，2018；Seo 等人，2018）。**基于检索的**方法将图像和地理坐标嵌入共享表示空间，并通过最近邻搜索进行定位。然而，这类方法依赖于规模庞大、分布全球且带有标注的图像数据库，而此类数据库的构建和维护成本高昂（Regmi 和 Shah，2019；Shi 等人，

2019，2020；Cepeda 等人，2023年）。近年来，**基于视觉语言模型**的方法利用大型多模态模型中编码的世界知识，直接预测位置，展现出良好的性能（Li 等人，2024年；Han 等人，2025年；Wang 等人，2025b）。相比之下，音频地理定位尚未得到探索。现有基准的多样性有限，往往聚焦于鸟类鸣叫等狭窄领域（Chasmai 等人，2025年）。

2.2 音频语言模型

音频语言模型的发展与深度学习的进步密切相关。早期的突破，例如 Deep Speech 2（Amodei 等人，2016年），以端到端神经网络取代了传统的多阶段流水线，实现了接近人类水平的转录性能。随后，自监督学习和弱监督学习的进一步发展推动了该领域的进步，Whisper（Radford 等人，2023年）便是其中的代表，它展现了可与监督方法相媲美的多语言转录性能。随着大模型预训练的兴起，音频已成为统一架构中的一等模态。近期的基础模型，如 GPT-4o（Achiam 等人，2023年）、Gemini3（Google DeepMind，2025年）和 Qwen3-Omni（Xu 等人，2025b），进一步体现了这一趋势。

目前已经提出了广泛的基准测试来评估音频语言模型，涵盖基础的语音到文本翻译（OpenAudioBench（Wang 等人，2025a））、人声声音分类（Vocal-Sound（Gong 等人，2022年）），以及空间推理（StarBench（Liu 等人，2025年））和因果发现（MECAT（Niu 等人，2025年））等更高层次的推理任务。然而，目前仍没有专门用于评估音频语言模型音频地理定位能力的基准测试。

3 音频地理定位基准测试

音频地理定位基准测试（AGL1K）旨在评估音频语言模型的地理组合推理能力。接下来，我们将介绍如何获取大规模音频-位置对（3.1）、执行粗略质量控制（3.2）、使用我们提出的音频可本地化性指标识别可可靠进行地理定位的录音（3.3），以及在获得音频语言模型的响应后进行后处理（3.4）。

3.1 数据集获取

全球范围的音频地理定位一直受到较少关注，主要原因是带有 GPS 标签的音频录音十分稀缺。我们发现了 Aporee[1]，，这是一个众包平台，用户可在其中上传全球各地带有地理标签的音频。我们与 Aporee 团队合作，获得了数万条音频-位置对，并以 Apache 许可证发布。有关 Aporee 的更多详细信息请参见附录 A.1。

3.2 初步筛选

由于 Aporee 录音由用户上传，其质量存在较大差异。为粗略移除低质量或信息量不足的样本，我们应用了四种针对常见失败原因的声学滤波器：低信号能量、类噪声内容、削波以及过于单调的录音。具体而言，我们计算 (1) 均方根能量 以排除振幅极低的音频，(2)频谱平坦度 以移除类噪声信号，(3)削波比率 以舍弃严重削波的录音，以及 (4)声学复杂度以消除时间变化有限的片段。这些滤波器保留了具有足够声学结构和潜在地理线索的录音。形式化定义见附录 A.3。

3.3 音频可定位性

3.3.1 可定位性计算

虽然初始声学滤波器会移除低质量录音，但无法确保剩余样本包含足够的地理信息量证据。为了构建输入可本地化的基准测试，我们提出了一种有理论依据的**音频可定位性**度量方法。

我们将音频录音的可定位性建模为信息丰富的（正向）与信息不足的（负向）声音类别所产生的净贡献。直观而言，接触信息丰富的类别（例如语言、具有地点特征的人类活动）的时间越长，地理可定位性就应越强；而信息不足的内容（例如婴儿哭声）持续越久，则越可能掩盖相关线索。不同类别的贡献也不尽相同：海岸声音或许能提供粗略证据，而语言内容往往能提供更精确的定位信息。基于此，我们将样本 k 的可定位性定义为

$$l_k = \sum_{i \in P} a_i t_{k,i} - \sum_{i \in N} \bar{a}_i t_{k,i}. \quad (1)$$

[1] https://aporee.org/maps/

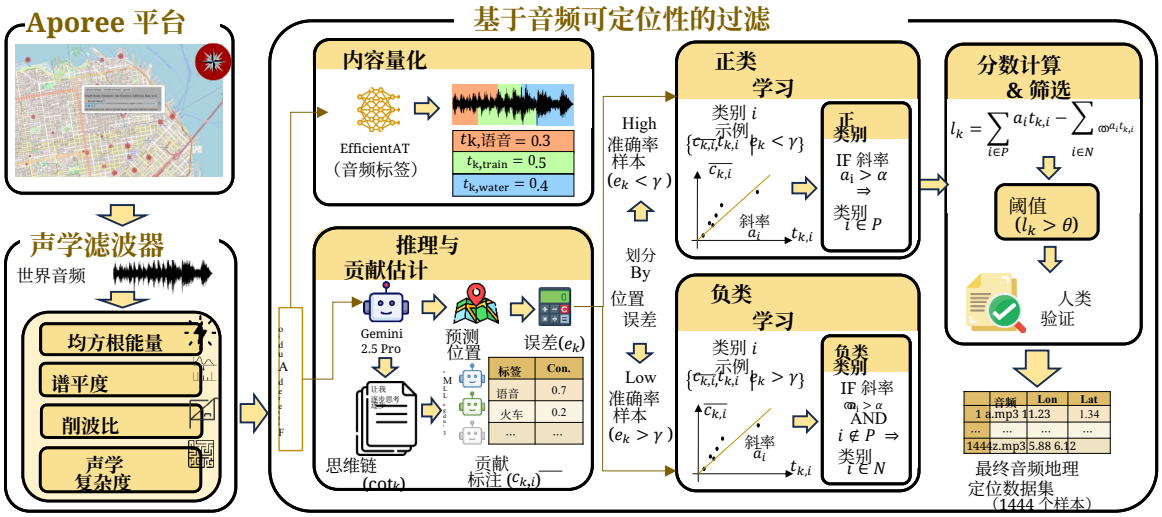


图2：**基准构建框架概览**。AGL1K 从众包平台 Aporee 平台整理而来。首先使用四种声学滤波器对录音进行筛选，随后采用我们提出的音频可定位性指标，该指标量化每个样本的地理信息量。

其中 P和 N 分别表示正向和负向声音类别的集合；$t_{k,\ i} \in [0,1]$表示类别 i 在样本 k中出现的时间比例；而 a_i （分别为 $\bar{a}_i$）表示类别 i 作为正向（分别为负向）类别时的贡献强度。我们通过使用 EfficientAT（Schmid等人，2023年）按照 AudioSet本体（Gem-meke等人，2017年）为每段录音添加标签，从而获得$t_{k,\ i}$。因此，估算类别层面的贡献是评估可定位性的关键。

由于音频地理定位缺乏成熟的专家启发式规则，我们根据强大的音频语言模型的推理行为推导类别贡献。对于每个样本 k，我们让 Gemini 2.5 生成预测位置和思维链 cot_k，并计算预测结果与真实坐标之间的距离误差 e_k。随后，我们提示三个语言模型基于cot_k，使用五级离散量表评估每个检测到的音频类别对模型推理的贡献。对三位评审的结果取平均，得到 $\bar{c}_{k,\ i}$，即类别 i对样本 k的估计贡献。

为确定哪个类别属于正类、哪个属于负类，我们对类别时长与评估贡献之间的关系进行简单线性拟合。对于包含 $\{k|e_k < \gamma\}$的样本，我们使用 $y = a_i x$ ，并采用$x = t_{k,\ i}$ 和$y = \bar{c}_{k,\ i}$进行拟合；若满足 $a_i > \alpha$，则将类别 i 纳入 P。对于包含 $\{k|e_k \geq \gamma\}$的样本，我们拟合 $y = \bar{a}_i x$ ，并将类别 i 在 N 满足 $\bar{a}_i > \alpha$ 且 $i \notin P$时。根据公式1，若录音满足 $l_k > \theta$，则认为其高度可定位。

应用上述标准后，我们得到一批高度可定位的录音。从中，我们人工筛选出 1,444 个高质量片段，其中包含人类语音和不含人类语音的样本数量均衡，构成用于音频地理定位基准测试的最终数据集。我们将 $\alpha = 1/3$ 、$\gamma =$ 设为 1000 km，并通过经验确定 $\theta = 1$。

3.3.2 最具代表性的正类和负类

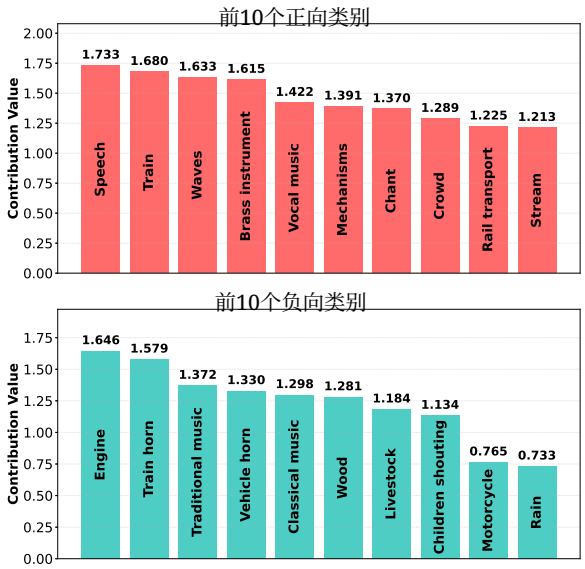


图3：**最具代表性的正类和负类。**

为评估所提出的可定位性指标的质量，我们列出了正向和负向贡献最大的 10 个声音类别（图 3）。由此可以看出几个明显的规律：语音是最突出的正向贡献因素，凸显了语言线索中蕴含的强烈地理特异性。此外，铁路运输和海浪等具有固有地域分布特征的声音也表现出较高的可定位性。相比之下，引擎或火车汽笛等通用噪声源，以及雨和木头声等全球普遍存在的自然声音，往往有显著的负向贡献，这可能是因为它们无处不在，且地理可辨识性有限。因此，**这些结果表明，学得的特征贡献与人类直觉高度吻合，支持了可定位性指标的可解释性和有效性**。

3.3.3 可定位性示例

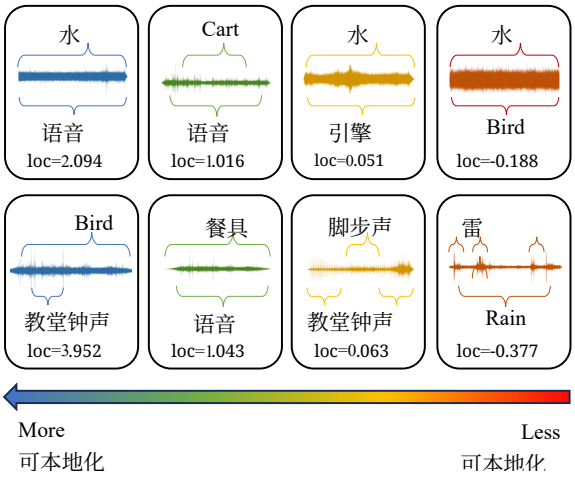


图4：**可定位性示例**。 可定位性从右向左递增。

图4 展示了八个示例片段计算出的可定位性。以雷声和雨声为主的录音得分最低，这可能是因为这些声音在世界大多数地区都很常见。虽然包含脚步声或教堂钟声的片段也不太容易定位，但鸟鸣的存在可以显著提高可定位性，因为物种特异性发声能提供更有力的地理线索。第二列中的片段包含语音，但语音被室内前景噪声（例如手推车和餐具）严重掩盖，因而地理信息量明显减弱。总体而言，**可定位性得分符合人类直觉，并支持我们将该指标作为筛选可本地化数据的可靠依据**。

3.4 后处理

按照上述流程，我们构建了 AGL1K 这一基准测试，其中包含广泛的音频类别，包括自然声音、动物发声、音乐和事物声音，以及许多融合多个类别的录音，从而提供丰富的地理信息线索。在地理范围上，该基准测试覆盖六大洲以及72个国家和地区以上。

在收集音频语言模型预测后，我们观察到，模型经常对同一区域使用别名或非正式变体（例如，将大洋洲报告为澳大利亚，或将美国报告为美国）。为确保评估的一致性，我们通过将此类变体映射到规范区域名称，将模型输出规范化为统一的地点体系，从而提高基准测试结果的可靠性。

此外，我们还开发了一个交互式音频地理定位网站，可部署在 Hugging Face Spaces 上，让全球用户探索并参与这项任务。请参阅附录 A.2。

4 实验

在本节中，我们探讨以下研究问题：**研究问题1**：当前的音频语言模型在所提出的基准测试中表现如何？**研究问题2**：音频语言模型如何根据声学线索推断潜在的地理位置？**RQ3**：音频语言模型在不同地区的地理位置预测中是否存在不平等现象？**RQ4**：音频地理定位中模型预测误差的主要原因是什么？

4.1 实验设置

4.1.1 基准测试模型

我们对 8 个最先进的闭源音频语言模型进行了基准测试，包括：GPT-4o Audio Preview（Achiam 等人，2023年）、Gemini 3 Pro（Google DeepMind，2025年）、Gemini 2.5 Pro（Comanici 等人，2025年）、Gemini 2.5 Flash、Gemini 2.5 Flash-Lite、Gemini 2.0 Flash-Thinking（Google，2025年）、Gemini 2.0 Flash 和 Gemini 2.0 Flash-Lite。我们还对 8 个高质量开源音频语言模型进行了基准测试，包括：Qwen3-Omni、Qwen2.5-Omni（Xu 等人，2025a）、Phi-4-MM1（Abouelenin 等人，2025年）、

表 1：不同模型的音频地理定位性能。闭源和开源模型组中的最佳结果分别以粗体标出，每组中的次佳结果则以下划线标出。

模型	Geo-分数↑	距离↓	洲际 准确率↑	国家 准确率↑	City 准确率↑	拒答率↓	<1 准确率↑	<10 acc.↑	<500 acc.↑	语音 dis.↓	非语音 dis.↓
随机	112.16	9869.01	0.14	0.01	0.00	0.00	0.00	0.00	0.00	9930.88	9717.65
闭源模型											
Gemini 3 Pro	**3031.82**	**2180.57**	**0.82**	**0.51**	**0.11**	0.01	**0.07**	**0.19**	**0.52**	**1548.29**	**3727.33**
Gemini 2.5 Pro	2826.95	2521.97	0.78	0.49	**0.11**	**0.00**	0.05	0.17	0.48	1634.22	4693.67
Gemini 2.5 Flash	2256.22	3558.37	0.65	0.39	0.07	0.01	0.02	0.12	0.36	2689.96	5682.74
Gemini 2.5 Flash-Lite	1687.97	4373.89	0.55	0.28	0.03	0.02	0.01	0.06	0.21	3742.97	5917.32
Gemini 2.0 Flash-Thinking	2494.46	2991.51	0.73	0.39	0.07	0.01	0.02	0.12	0.38	2265.47	4767.61
Gemini 2.0 Flash	2535.49	2906.31	0.73	0.40	0.08	0.01	0.02	0.13	0.41	2189.30	4660.31
Gemini 2.0 Flash-Lite	2417.14	3223.85	0.71	0.38	0.06	0.04	0.01	0.11	0.37	2296.28	5492.96
GPT-4o Audio Preview	1996.15	4067.87	0.61	0.37	0.05	0.04	0.02	0.10	0.30	2841.03	7069.09
开源模型											
Qwen3-Omni	**1498.92**	5174.36	0.47	0.25	0.02	0.04	**0.01**	**0.05**	**0.20**	4741.73	6235.23
Qwen2.5-Omni	1317.74	5476.83	0.45	**0.26**	0.02	0.31	0.00	0.04	0.17	4540.06	7777.20
Phi-4-MM1	826.88	6462.43	0.33	0.08	0.01	0.15	0.00	0.01	0.08	6311.65	6831.08
Kimi-Audio	1336.20	5590.20	0.43	0.22	0.02	0.37	0.00	0.04	0.18	4634.32	7928.59
Gemma-3n-E4B-it	1161.54	5815.46	0.41	0.17	0.01	0.05	0.00	0.04	0.14	5402.73	6825.13
MiniCPM-o-2.6	1080.05	6600.83	0.44	0.22	0.02	0.54	0.00	0.01	0.14	5572.78	9115.74
Mimo-Audio	1445.39	**4853.25**	**0.54**	0.20	**0.03**	0.11	0.00	0.03	0.16	**4113.93**	6661.84
Mimo-Audio-Think	1447.30	5008.01	0.51	0.20	**0.03**	**0.02**	0.00	0.04	0.18	4599.73	**6006.78**

Kimi-Audio（KimiTeam 等人，2025年）、Gemma-3n-E4B-it（Team 等人，2025年）、MiniCPM-o-2.6（Yao 等人，2024年）、Mimo-audio（小米，2025年）和 Mimo-audio-think（小米，2025年）。对于除 Qwen3-Omni 外的开源模型，我们将其部署在 RTX 4090 上。其他模型则通过 API 并使用默认设置进行访问。

4.1.2 评估指标

评估指标包括平均距离误差（公里）δ、基于 GeoGuessr 游戏定义的地理得分，其定义为 $5000\exp(-\delta/1492.7)$，分层的大洲/国家/城市级准确率，以及 1 公里、10 公里和 500 公里阈值下的准确率（记作 $<*$），用于衡量细粒度和粗粒度的定位性能。我们还报告模型的拒答率（未作答的比例），作为稳健性的指标。最后，为了分析对语音线索的依赖程度，我们分别计算语音音频片段和非语音音频片段的平均距离误差。各项指标的定义见附录A.4。

4.2 比较结果（研究问题 1）

现代音频语言模型已开始展现音频地理定位方面的新兴能力。表格2 报告了所有模型在音频地理定位基准测试中的表现。结果显示，Gemini 3 Pro 的整体表现最为出色，19% 的示例定位在 10 公里范围内；大陆级准确率为 0.82，国家级准确率为 0.51。然而，闭源模型与开源模型之间存在明显差距：表现最好的开源模型 Mimo-audio 的平均误差为 4853 公里，是 Gemini 3 Pro 的两倍多。

音频地理定位性能随着模型容量的增加而呈现系统性提升。在 Gemini 系列中，我们观察到明显的进步。距离误差从 Gemini 2.0 Flash-Thinking 的 2992 公里降至 Gemini 3 Pro 的 2181 公里，各项准确率也稳步提升。值得注意的是，Gemini 2.0 Flash 的一些变体表现优于 Gemini 2.5 Flash，这表明蒸馏模型可能优先考虑效率或特定任务目标，而这些目标与音频地理定位并不直接一致。然而，增强推理变体（例如“思考”模式）并未带来稳定的提升，这说明单靠显式推理的益处有限。

语言内容是基于音频进行地理定位的主要支架。对于 Gemini 3 Pro，存在语音时，平均定位误差从非语音音频中的 3727 千米降至 **1548 千米**。从闭源模型到开源模型，所有评估的音频语言模型中都出现了类似差距。这一结果凸显了语言线索的主导作用，同时也表明，纯粹的环境声音仍然更具挑战性。

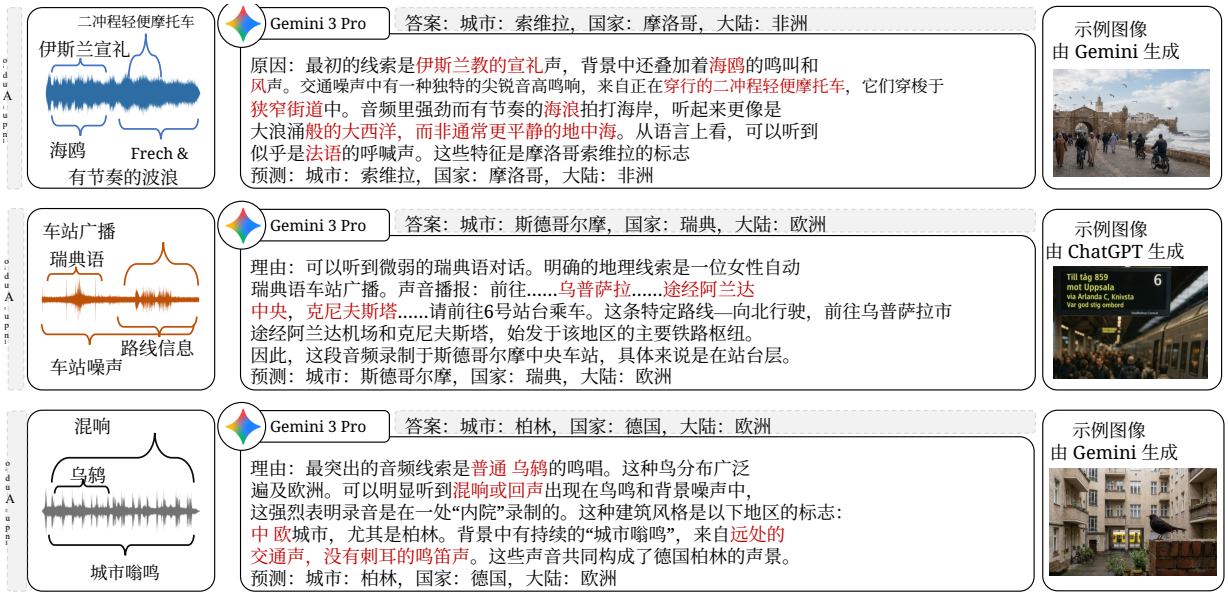


图5：**基　准示例**。我们选取三个具有代表性的音频样本，并展示其音频线索的分布，它们的音频线索中包括 Gemini 3 Pro 的推理过程，以及另外三个音频语言模型的预测。

更多统计数据和比较结果见附录A.6。

4.3 基准示例（研究问题2）

为定性考察现代音频模型的推理行为，我们分析了 Gemini 3 Pro 在三个具有代表性的场景中的推理输出，如图 5 所示，同时展示 GPT-4o Audio、Qwen3-Omni 和 Mimo-Audio 的预测。完整的推理轨迹见附录 B.6。

第一个案例需要整合语言和环境信号。除了识别伊斯兰宣礼外，Gemini 3 Pro 还检测到海岸风声、海鸥叫声、法语语音以及狭窄街道中的摩托车噪声，并综合判断录音地点为摩洛哥索维拉。然而，大多数其他模型过度强调宣礼，因此将地点预测为伊斯坦布尔或孟买。

第二个案例考查如何根据间接线索进行地理推断。音频提到经由阿兰达和克尼夫斯塔前往乌普萨拉，借助了区域交通知识。Gemini 3 Pro、GPT-4o 和 Mimo-Audio 均正确推断录音地点为斯德哥尔摩。

这第三个案例中没有口语。相反，Gemini 3 Pro 依靠非语言线索，包括乌鸫的叫声、建筑风格和背景交通模式，正确地将场景定位到德国柏林。然而，Qwen3 和 Mimo 认出了这只欧洲鸟类，未经进一步思考便随机猜了一个欧洲的地点。

总的来说，这些案例研究表明，**领先的音频模型能够利用语言、环境和语境线索进行地理推理**，而且**我们的基准测试通过统一的评估框架捕捉了这些组合推理模式**。

4.4 洲级预测不平等（RQ3）

图6展示了洲级混淆情况（按行归一化）。Gemini 3 Pro 的对角线最明显，说明其大陆识别最稳定，尤其是在非洲、亚洲和欧洲；但在大洋洲和南美洲的表现有所下降。GPT-4o 音频的对角线分布较不集中，跨大陆误判更为普遍，其中大洋洲尤其不稳定，常被预测为其他大洲。通义千问Qwen3-Omni 显示出明显的标签坍塌倾向，偏向北美洲：许多来自其他大洲的样本都被预测为北美洲，导致其在大洋洲和南美洲的表现不佳。总体而言，这些结果表明，**地理预测不平等是音频语言模型普遍存在的问题，这凸显了未来研究需要明确考虑区域公平性**。

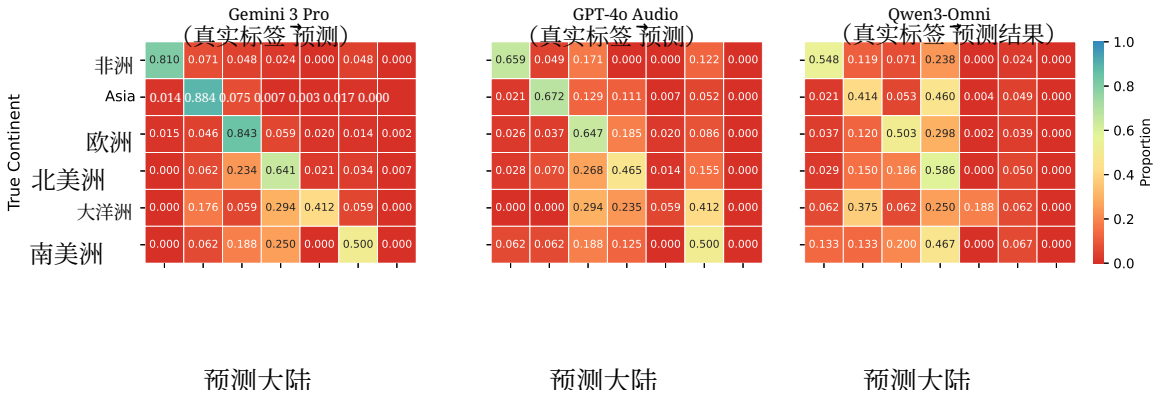


图6：音频地理定位中的洲级预测不平等。第 i 行、第 j 列的条目表示真实标签属于第 i 行所示大陆的部分区域，却被预测为第 j 行所示大陆。

4.5 错误分布（RQ4）

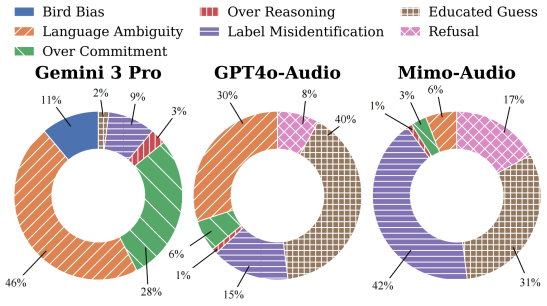


图7：三个模型的错误分布。

为分析模型失效原因，我们对三种具有代表性的模型在国家级误分类中的300个样本进行了标注。误差分为七类：鸟类偏差、语言歧义、过度承诺、过度推理、标签误识别、有根据的猜测和拒答，涵盖感知和推理相关失误。附录中进一步阐述了各类误差的定义B。

图 7 显示，不同模型的错误分布存在显著差异。对于 Gemini 3 Pro，语言歧义占主导，其次是过度承诺和鸟类偏差。这反映了模型倾向于利用细粒度的声学线索，尤其是语言和鸟类鸣叫；这提高了定位准确率，但也会带来独特的错误。相比之下，GPT-4o Audio 主要受到有根据的猜测和语言歧义的影响，表明当有效线索不足时，模型经常退而采用不确定的预测。Mimo-Audio 则呈现出不同的模式，其中，标签误识别占错误总数的最大比例，其次是有根据的猜测和拒答，这表明模型在多语言音频辨别和置信度校准方面存在局限。

总之，这项实验指出了若干切实可行的方向，可用于改进不同模型系列的音频地理定位。在感知层面，加强多语种语音识别对于减少语言相关错误至关重要。在推理层面，模型应采用更审慎且注重证据的推理方式，以尽量减少对缺乏依据的有根据的猜测的依赖。在融合层面，模型应避免过度依赖语言或鸟类鸣叫等单一线索，而应整合多种互补的音频信号。总之，**这些发现为提升基于音频的定位系统的准确率和鲁棒性提供了实用指导**。

5 结论

我们推出 AGL1K，这是一个用于音频地理定位的基准测试。该基准测试基于用户上传的录音，并按照我们提出的音频可定位性标准进行筛选，包含来自六大洲72个国家和地区的1,444个音频片段，涵盖多样的声学场景。通过评估16种音频语言模型，我们发现音频语言模型已展现出地理定位能力，但 AGL1K 仍具有挑战性。我们的分析识别出主要错误模式，并提出了改进音频语言模型的方向，包括增强细粒度感知、减少区域偏差以及提升组合推理能力。由于音频地理定位需要基于音频线索进行组合推理，

我们预计，音频地理定位将成为评估音频语言模型推理能力的重要基准测试。

6 局限性

- 由于原始数据本身的分布特点，各大洲的样本数量并不均衡。尤其是，与非洲、大洋洲和南美洲等其他地区相比，欧洲、亚洲和北美洲的样本数量明显过多。

- 声音标签的多样性，以及正向和负向归因结果的分布，在一定程度上受到所选标注模型和所收集数据的限制。

参考文献

AbdelrahmanAbouelenin、Atabak Ashfaq、Adam Atkin- son、Hany Awadalla、Nguyen Bach、Jian min Bao、Alon Benhaim、Martin Cai、Vishrav Chaudhary、Con- gcong Chen 等人。2025年。Phi-4-mini 技术报告：通过 LoRA 混合实现紧凑而强大的多模态语言模型。*arXiv* 预印本 *arXiv:2503.01743*。

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等人。2023年。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*。

Dario Amodei、Sundaram Ananthanarayanan、Rishita Anubhai、Jingliang Bai、Eric Battenberg、Carl Case、Jared Casper、Bryan Catanzaro、Jingdong Chen、Mike Chrzanowski、Adam Coates、Greg Diamos、Erich Elsen、Jesse H. Engel、Linxi Fan、Christopher Fougner、Awni Y. Hannun、Billy Jun、Tony Han及其他20位作者。2016年。Deep Speech 2：端到端英语和普通话语音识别。载于第*33*届国际机器学习会议论文集，*ICML 2016*，纽约市，美国纽约州，*2016*年*6*月*19*日至*24*日，第48卷，收录于*JMLR*研讨会与会议论文集，第173–182页。JMLR.org。

Vicente Vivanco Cepeda、Gaurav Kumar Nayak和Mubarak Shah。2023年。GeoCLIP：受CLIP启发的位置与图像对齐方法，用于实现有效的全球地理定位。载于神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统大会年会论文集，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日。

Mustafa Chasmai、Wuao Liu、Subhransu Maji 和 Grant Van Horn。2025年。音频地理定位：自然声音基准。*CoRR*，abs/2505.18726。

Brandon Clark、Alec Kerrigan、Parth Parag Kulkarni、Vicente Vivanco Cepeda 和 Mubarak Shah。2023年。我们身在何处，又在看什么：利用层级结构和场景进行基于查询的全球图像地理定位。载于*IEEE/CVF* 计算机视觉与模式识别会议，*CVPR 2023*，温哥华，不列颠哥伦比亚省，加拿大，*2023*年*6*月*17*日至*24*日，第23182–23190页。IEEE。

Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Mar- cel Blistein、Ori Ram、Dan Zhang、Evan Rosen及其他1位作者。2025年。Gemini 2.5：凭借高级推理、多模态、长上下文和新一代智能体能力，推动前沿发展。*arXiv*预印本 *arXiv:2507.06261*。

Jort F. Gemmeke、Daniel P. W. Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R. Channing Moore、Manoj Plakal和Marvin Ritter。2017年。Audio set：用于音频事件的本体和人工标注数据集。载于*2017*年 *IEEE*国际声学、语音与信号处理会议（*ICASSP 2017*），美国路易斯安那州新奥尔良，*2017*年*3*月*5*日至*9*日，第776–780页。IEEE。

Yuan Gong、Jin Yu 和 James R. Glass。2022年。Vocal-sound：用于提升人类发声识别的数据集。发表于*IEEE*国际声学、语音与信号处理会议（*ICASSP 2022*），线上及新加坡，*2022*年*5*月*23*日至*27*日，第151–155页。IEEE。

Google。2025年。Gemini 2.0 Flash：模型卡。模型卡（技术报告），Google。模型卡发布于2025年4月15日。

Google DeepMind。2025年。Gemini 3 Pro：模型卡。模型卡（技术报告），Google DeepMind。模型卡发布于2025年11月。

Xiao Han、Chen Zhu、Hengshu Zhu 和 Xiangyu Zhao。2025年。地理定位中的群体智能：多智能体大型视觉语言模型协作框架。发表于第*31*届*ACM SIGKDD*知识发现与数据挖掘会议论文集，第*2*卷，*KDD 2025*，加拿大安大略省多伦多，*2025*年*8*月*3*日至7日，第814–825页。ACM。

KimiTeam、Ding Ding、Zeqian Ju、Yichong Leng、Songxiang Liu、Tong Liu、Zeyu Shang、Kai Shen、Wei Song、Xu Tan、Heyi Tang、Zhengtao Wang、Chu Wei、Yifei Xin、Xinran Xu、JianweiYu、Yutao Zhang、Xinyu Zhou、Y. Charles 等 21 人。2025年。Kimi- 音频技术报告。预印本，arXiv:2504.18425。

Ling Li、Yu Ye、Bingchuan Jiang 和 Wei Zeng。2024年。Georeasoner：利用大型视觉语言模型对街景进行推理以实现地理定位。发表于第四十一届国际机器学习大会，*ICML 2024*，维也纳，奥地利，*2024*年*7*月*21*日至*27*日。OpenReview.net。

Zihan Liu、Zhikang Niu、Qiuyang Xiao、Zhisheng Zheng、Ruoqi Yuan、Yuhang Zang、Yuhang Cao、Xiaoyi Dong、Jianze Liang、Xie Chen、LeileiSun、

DahuaLin，和 Jiaqi Wang。2025年。 Star-bench：探究音频 4D 智能中的深层时空推理。 *CoRR*，abs/2510.24693。

Eric Müller-Budack、Kader Pustu-Iren 和 Ralph Ew- erth。2018。使用层次模型和场景分类进行照片地理位置估计。载于 计算机视觉——*ECCV 2018*——第*15*届欧洲计算机视觉会议，慕尼黑，德国，*2018*年*9*月*8*日至*14*日，论文集，第*XII*部分，第11216卷，收录于《计算机科学讲义》，第575–592页。施普林格。

Yadong Niu、Tianzi Wang、Heinrich Dinkel、Xingwei Sun、Jiahao Zhou、Gang Li、Jizhong Liu、Xunying Liu、Junbo Zhang 和 Jian Luan。2025年。MECAT：一个由多位专家构建的细粒度音频理解任务基准测试。*CoRR*，abs/2507.23511。

Alec Radford、Jong Wook Kim、Tao Xu、Greg Brock- man、Christine McLeavey 和 Ilya Sutskever。2023年。通过大规模弱监督实现稳健的语音识别。载于国际机器学习会议，*ICML 2023*，*2023*年7月*23*日至*29*日，檀香山，夏威夷，美国，第202卷，收录于机器学习研究论文集，第28492–28518页。PMLR。

Krishna Regmi 和 Mubarak Shah。2019。弥合地面与航拍图像匹配的领域差距。载于*2019 IEEE/CVF*计算机视觉国际会议，*ICCV 2019*，韩国首尔，*2019*年*10*月*27*日至*11*月*2*日，第470–479页。IEEE。

Florian Schmid、Khaled Koutini 和 Gerhard Widmer。2023年。通过Transformer到CNN知识蒸馏实现高效的大规模音频标签。载于*IEEE*国际声学、语音与信号处理会议 *ICASSP 2023*，希腊罗得岛，*2023*年*6*月*4*日至*10*日，第1–5页。IEEE。

Paul Hongsuck Seo、Tobias Weyand、Jack Sim 和 Bo-hyung Han。2018。Cplanet：通过地图组合划分增强图像地理定位。载于计算机视觉——*ECCV 2018*——第*15*届欧洲会议，德国慕尼黑，*2018*年*9*月*8*日至*14*日，论文集，第*X*部分，第11214卷，《计算机科学讲义》，第544–560页。施普林格。

Yujiao Shi、Liu Liu、Xin Yu 和 Hongdong Li。2019。用于图像跨视角地理定位的空间感知特征聚合。载于《神经信息处理系统进展》第*32*卷：*2019*年神经信息处理系统年度会议，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，第10090–10100页。

Yujiao Shi、Xin Yu、Dylan Campbell 和 Hongdong Li。2020。我正在看哪里？通过跨视角匹配进行联合位置与方向估计。载于*2020*年*IEEE/CVF*计算机视觉与模式识别会议，*CVPR 2020*，美国华盛顿州西雅图，*2020*年*6*月*13*日至*19*日，第4063–4071页。计算机视觉基金会 / IEEE。

Gemma团队、Aishwarya Kamath、Johan Ferret、Shreya Pathak、Nino Vieillard、RamonaMerhej、Sarah Perrin、Tatiana Matejovicova、Alexandre Ramé、Morgane Rivière 和另外 1 位作者。2025年。Gemma 3 技术报告。*arXiv*预印本*arXiv:2503.19786*。

Bin Wang、Xunlong Zou、Geyu Lin、Shuo Sun、Zhuohan Liu、Wenyu Zhang、Zhengyuan Liu、AiTi Aw 和 Nancy F. Chen。2025a。AudioBench：音频大语言模型通用基准测试。载于《*2025*年美洲计算语言学协会分会会议论文集：人类语言技术》，*NAACL 2025*——第*1*卷：长篇论文，美国新墨西哥州阿尔伯克基，*2025*年*4*月*29*日至*5*月*4*日，第4297–4316页。计算语言学协会。

Chun Wang、Xiaoran Pan、Zihao Pan、Haofan Wang 和 Yiren Song。2025b。GREsuite：通过微调视觉语言模型和增强推理链进行地理定位推理。*CoRR*，abs/2505.18700。

Tobias Weyand、Ilya Kostrikov 和 James Philbin。2016年。Planet——使用卷积神经网络进行照片地理定位。载于计算机视觉——*ECCV 2016*——第*14*届欧洲会议，阿姆斯特丹，荷兰，*2016*年*10*月*11*日至*14*日，论文集，第*VIII*部分，《计算机科学讲义》第卷*9912*，第37–55页。施普林格。

小米LLM-Core团队。2025年。Mimo-Audio：音频语言模型是少样本学习器。

Jin Xu、Zhifang Guo、Jinzheng He、Hangrui Hu、Ting He、Shuai Bai、Keqin Chen、Jialin Wang、Yang Fan、Kai Dang 等人。2025a。Qwen2.5-omni 技术报告。*arXiv* 预印本 *arXiv:2503.20215*。

Jin Xu、Zhifang Guo、Hangrui Hu、Yunfei Chu、Xiong Wang、Jinzheng He、Yuxuan Wang、Xian Shi、Ting He、Xinfa Zhu、Yuanjun Lv、Yongqi Wang、Dake Guo、He Wang、Linhan Ma、Pei Zhang、Xinyu Zhang、Hongkun Hao、Zishan Guo及其他19位作者。2025b。通义千问3-omni 技术报告。预印本，arXiv:2509.17765。

Yuan Yao、Tianyu Yu、Ao Zhang、ChongyiWang、Junbo Cui、Hongji Zhu、Tianchi Cai、Haoyu Li、Weilin Zhao、Zhihui He 等人。2024年。MiniCPM-V：媲美 GPT-4V 的手机端多模态大语言模型。*arXiv* 预印本 *arXiv:2408.01800*。

A 附录

A.1 Aporee 平台简介

Aporee 于 2006年推出，用户可自愿上传带有地理标签的环境声音录音，并附上 GPS 坐标；平台由此构建起全球声音地图。该平台涵盖各种各样的声学环境，包括城市地区、农村地区和自然景观，广泛用于艺术、教育和研究。通过与 Aporee 团队合作，我们获得了在 Apache 许可证下使用其音频数据的权限。

A.2 交互式平台构建

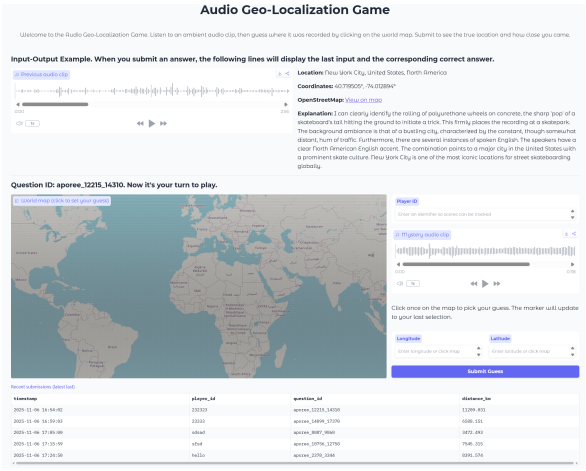


图 8：**我们的交互式平台截图**。该平台由 Hugging Face Gradio 构建，可轻松部署到 Hugging Face Space 或本地。

为了收集人类表现数据并鼓励更多人参与音频地理定位，我们开发了一个交互式网络平台，并将其部署在 HuggingFace 上。界面向用户呈现一个音频片段和一张交互式地图，用户需要点击地图来推断声音的位置。用户的预测结果和匿名标识符会公开记录在 HuggingFace Space 上，从而能够分析人类定位行为，并为比较模型表现提供参考基线。

A.3 四种声学滤波器的形式化定义

四个声学滤波器定义如下：

$$\begin{aligned}
\mathrm{RMS} &= \sqrt{\frac{1}{N}\sum_{n=1}^{N} x[n]^2}, \\
\mathrm{SF} &= \frac{\left(\prod_{k=1}^{K} P[k]\right)^{1/K}}{\frac{1}{K}\sum_{k=1}^{K} P[k]}, \\
\mathrm{CR} &= \frac{\#\{n \mid |x[n]| = x_{\max}\}}{N}, \\
\mathrm{AC} &= \sum_{f} \frac{\sum_{t=1}^{T-1} |E_{f,t+1} - E_{f,t}|}{\sum_{t=1}^{T} E_{f,t}},
\end{aligned} \tag{2}$$

其中， $x[n]$ 表示样本 n处的音频振幅，$P[k]$ 表示频率箱 k处的功率谱密度，$E_{f,\ t}$ 表示频带 f 在时间帧 t处的能量。

A.4 评估指标的定义

设 $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$ 表示评估数据集，其中 x_i 是一个音频样本， $y_i = (c_i^{\text{cont}}, c_i^{\text{country}}, c_i^{\text{city}}, \ell_i)$表示其真实的大洲、国家、城市和地理坐标 $\ell_i = (\phi_i, \lambda_i)$（纬度和经度）。对于模型预测 $\hat{}\ y_i = (\hat{}\ c_i^{\text{cont}}, \hat{}\ c_i^{\text{country}}, \hat{}\ c_i^{\text{city}}, \hat{}\ \ell_i)$，我们定义以下指标。

A.4.1 平均距离误差（km）

我们使用半正矢公式计算预测坐标与真实坐标之间的大圆距离：

$$d_i = \mathrm{Haversine}(\ell_i, \hat{\ell} * i). \tag{3}$$

如果模型拒绝作答，我们直接将 d_i设为10000。平均距离误差为

$$\mathrm{DistErr} = \frac{1}{|\mathcal{A}|}\sum * i \in \mathcal{A} d_i, \tag{4}$$

其中 $\mathcal{A} \subseteq 1, \ldots, N$表示所有样本的集合。

A.4.2 层次准确率

我们报告三个地理层级的准确率：

$$\begin{aligned}
\mathrm{Acc}_L &= \frac{1}{|\mathcal{A}|}\sum_{i \in \mathcal{A}} \mathbb{I}\left[\hat{c}_i^{L} = c_i^{L}\right], \\
L &\in \text{continent}, \text{country}, \text{city},
\end{aligned} \tag{5}$$

其中 $\mathbb{I}[\cdot]$ 是指示函数。

表2：进一步比较结果。

模型	q25↓	q50↓	q75↓	动物距离↓	音乐距离↓	自然距离↓	事物距离↓
随机	6253.81	10062.38	13639.9	10181.77	10667.35	10259.51	10515.23
闭源模型							
Gemini 3 Pro	**55**	**480.19**	**1888.6**	**3143.16**	**3231.5**	**3669.51**	**3765.42**
Gemini 2.5 Pro	97.59	548.1	2848.37	4635.67	3476.44	4620.69	4541.46
Gemini 2.5 Flash	279.21	1213.16	6782.78	5962.53	3893.8	5918.76	5353.91
Gemini 2.5 Flash-Lite	664.15	2295.73	8275.64	5976.78	4092.25	6247.15	5911.77
Gemini 2.0 Flash Thinking	263.39	883.79	5544.65	4576.98	4889.07	4839.28	4808.03
Gemini 2.0 Flash	218.81	886.77	5119.63	4510.66	4319.07	4673.66	4462.81
Gemini 2.0 Flash-Lite	280.3	1002.83	5888.17	5936.17	3900.27	5883.09	5426.24
GPT-4o Audio Preview	415.21	1715.7	7723.63	7960.98	4849.44	7696.62	6613.44
开源模型							
Qwen3-Omni	**633.98**	5627.17	**8821.64**	**5920.09**	5048.03	6064.08	6484.72
Qwen2.5-Omni	983.09	5952.77	10000	7723.77	5694.41	7871.21	7854.18
Phi-4-MM1	1797.82	6485.13	10000	6905.37	6066.34	6728.47	6640.15
Kimi-Audio	829.04	5945.86	10000	8274.65	6692.15	8148.38	7601.32
gemma-3n-E4B-it	1263.42	7020.47	9441.25	7203.56	6063.93	7160.14	6794.96
MiniCPM-o-2.6	1355.05	10000	10000	9517.41	7424.07	9345.83	9070.55
Mimo-Audio	893.83	**3450.54**	9208.35	7318.02	4552	7293.23	6091.1
Mimo-Audio-think	854.09	3627.95	9036.6	6046.2	**4419.81**	**6028.08**	**6001.72**

A.4.3 阈值距离准确率（$< \tau$）

为评估由粗到细的定位性能，我们在多个距离阈值下计算阈值准确率 $\tau \in 1$，10，100，1000 千米：

$$\mathrm{Acc}_{<\tau} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathbb{I}\big[d_i < \tau\big]. \tag{6}$$

这些对应于以下列：< 1，< 10，< 100，以及 < 1000表格2。

A.4.4 拒答率

某些模型可能会选择不作答。我们将拒答率定义为

$$\mathrm{RejectRate} = 1 - \frac{|\mathcal{A}_a|}{N}, \tag{7}$$

它衡量模型未给出预测的样本所占的比例，而 $\mathcal{A}_a$ 表示模型给出有效预测（即未拒答）的样本集合。

A.4.5 语音与非语音距离误差

为分析对语言线索的依赖，我们将数据集划分为语音样本 $\mathcal{D}$ *语音和非语音样本 $\mathcal{D}_{\text{non-speech}}$。随后，我们分别计算平均距离误差：

$$\begin{aligned} \mathrm{DistErr}_{\text{speech}} &= \frac{1}{|\mathcal{A}_{\text{speech}}|} \sum_{i \in \mathcal{A}_{\text{speech}}} d_i, \\ \mathrm{DistErr}_{\text{non-speech}} &= \frac{1}{|\mathcal{A}_{\text{non-speech}}|} \sum_{i \in \mathcal{A}_{\text{non-speech}}} d_i. \end{aligned} \tag{8}$$

A.5 主要声音标签的定义

自然指的是 Au- dioset 中的自然声音标签。其定义为：由自然来源在其正常声景中产生的声音，不包括动物和人类的声音。

动物指 AudioSet 中的“动物”标签。其定义为：由非人类动物的身体和行为产生的所有声音。

音乐指 AudioSet 中的“音乐”标签。其定义为：音乐是一种艺术形式和文化活动，其媒介是声音与静默。音乐的常见要素包括音高、节奏、力度，以及音色和织体等声音特质。

事物 指 AudioSet 中的“事物声音”标签。其定义为：一组声音类别，指听众能够立即理解为源自特定物体的声音（而非更字面地听作“声音”）。

A.6 更多比较结果

我们在表2中报告了每个模型在第 25、50 和 75 个百分位数上的定位误差（分别记为 q25、q50 和 q75）。此外，我们还将误差细分为四个子类别：动物、音乐、自然和事物声音。值得注意的是，Gemini 3 Pro 在所有类别的第 25 个百分位数上的定位误差均低于 50 千米，凸显了其卓越的音频定位能力。相比之下，表现最佳的开源

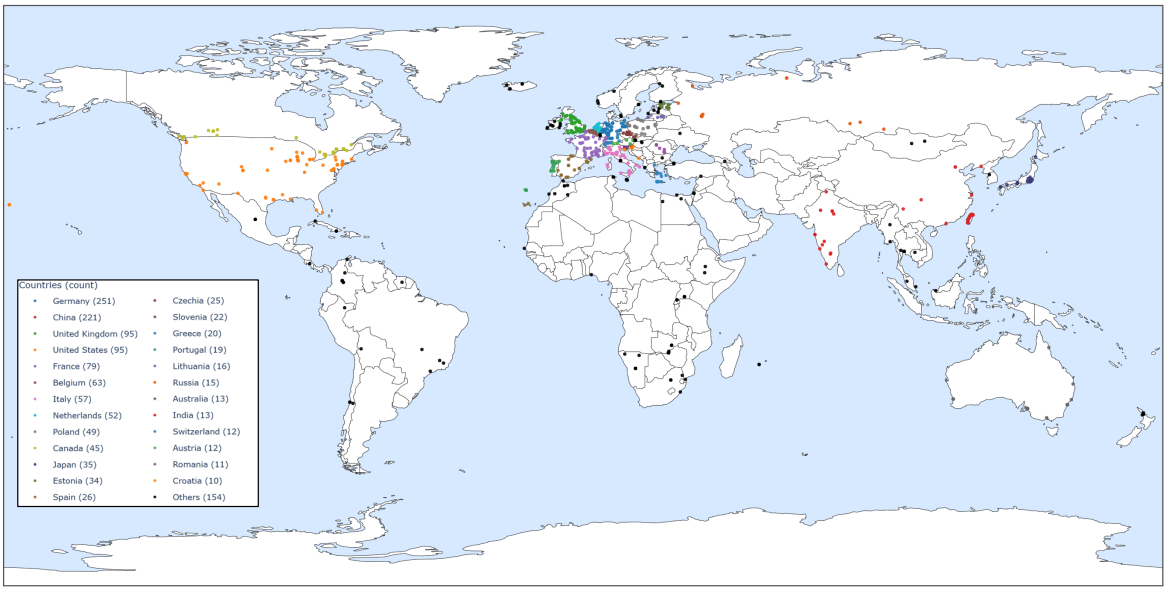


图 9：**Th AGL1K 的全球分布**。 74 个国家/地区的 1,444 段音频的位置。每个标记表示一个图像坐标，颜色表示国家/地区（列出贡献最多的国家/地区；“其他”汇总其余部分 er)

模型在相同百分位数上的误差仍达到 634 千米，突显了专有模型与开源模型之间显著的性能差距。音乐和动物类别中的误差相对较低，可能是因为音乐内容和动物发声中存在语言或分布线索；而自然和事物声音类别通常缺乏此类信息结构。

A.7 更多实验设置

我们通过 OpenRouter（ https://openrouter.ai/models ）访问闭源模型，并通过阿里云的百炼 API 调用 Qwen3-Omni（ https://cn.aliyun.com/product/bailian ）。定价详情请参阅相应网站。开源模型部署在单张 NVIDIA RTX 4090 GPU 上，每个模型的推理约需 12 GPU 小时。

B 各类误差的定义

鸟类偏差指由特定地区的鸟类鸣叫引起的误差。

语言歧义指某种语言在多个地理区域通用，导致模型虽然正确识别出该语言，却选择了错误位置的情况。

过度承诺 描述的是模型的中间推理支持多种个合理假设，但它仍然做出了单一且过度自信的预测。

过度推理描述模型最初推断出正确答案，却因随后进行不必要或错误引导的额外推理而推翻该答案的情况。标签误识别对应事实推断错误，最常见的是语言识别错误（例如将法语归类为德语）。

有根据的猜测指模型表示信息不足，实际上随机作答的情况。

拒答涵盖模型拒绝提供答案的情况。

B.1 AGL1K 的全球分布

图9 展示了 AGL1K 的全球分布。

B.2 三个贡献标注大语言模型的一致性

为评估大语言模型在贡献评估中的可靠性和遵循指令的一致性，我们使用三项互补指标衡量 Gemini 、 Claude 和通义千问之间的模型间一致性（图 10）。在样本层面，贡献向量的两两余弦相似度介于 **0.65 至 0.68**，表明对于各个示例中哪些音频线索至关重要，各模型在方向上高度一致。当汇

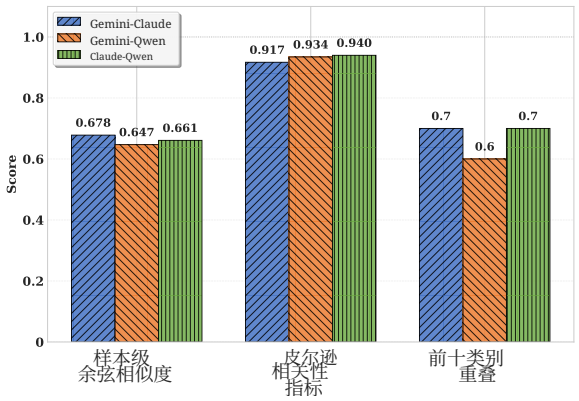


图10：不同指标下的模型相似性比较。

总各类别的贡献时，所有模型对之间的皮尔逊相关系数均超过**0.91**，显示出高度一致的全局归因模式。此外，前十类别重叠度始终高于**0.60**，说明各模型对大多数最具影响力的声音类别看法一致。

总体而言，这些结果表明，不同大语言模型对贡献的判断具有稳定性，支持我们基于集成的标注策略具有鲁棒性，同时也允许模型存在细微的特定偏好。

B.3 可定位性的洲际分布

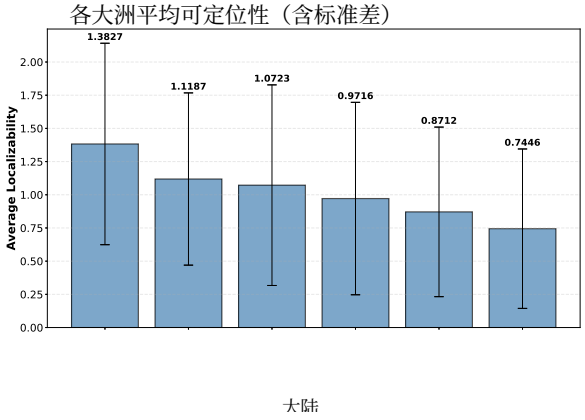


图11：可定位性的洲际分布。

我们分析了音频在不同大洲的可定位性，如图 11所示。我们观察到，非洲、亚洲和南美洲的平均可定位性得分较高，而欧洲、北美洲和大洋洲的得分较低。一种可能的解释在于前一组地区的语言和文化特征更为鲜明：非洲、亚洲和南美洲的地区通常具有更独特的声学环境和语言，这可能为地理定位提供更丰富的线索。相比之下，欧洲、北美洲和大洋洲的国家通常以英语为主导语言，因此音频模式的区分度可能较低，进而导致可定位性下降。

此外，观察到的可定位性排名与模型的实际表现部分一致。例如，Gemini 3 Pro 在亚洲的预测准确率最高，其次是欧洲、非洲、北美洲、南美洲和大洋洲。这与各大洲的平均可定位性得分相关。这表明，可定位性这一指标捕捉到了会影响模型表现的音频内在属性，并且可作为衡量不同地区地理定位难度的有效预测指标。

B.4 人类标注员

本研究在两个关键阶段邀请了人类标注员参与。首先，在数据集优化阶段，根据可定位性指标筛选出约 3,000 个样本后，一名人类标注员负责挑选出确实可进行地理定位的示例。这名计算机科学专业的博士生被要求综合考虑模型的输出、音频内容和真实位置，判断是否有可能实现城市级推断。

其次，在错误分析阶段，另一名人类标注员负责根据预先定义的错误类别对失败案例进行标注。对于语音相关错误，标注员会咨询一名语言学专业的博士生，以判断语言识别是否有误。标注员的报酬为每小时 100 人民币，并获知其标注结果将用于本研究。

B.5 提示词

B.5.1 基准测试提示词

基准测试提示词

你是一位音频分析师，擅长根据环境声音识别地理位置。请仔细聆听提供的音频录音，并识别录音的地理位置

录制地点。
无论是否确定，你都必须给出自己最好的猜测。即使你觉得不确定，也仍须根据音频特征、环境声音、所说的语言、文化指标或你能检测到的任何其他音频线索，做出有根据的猜测。你的回答必须采用以下 JSON 格式（将值替换为你的预测）：

{ "reason": "详细说明你的推理过程。描述你听到的音频线索（例如语言、口音、背景声音、环境噪音、文化指标、天气声音、交通模式、动物声音等），以及这些线索如何引导你做出这一具体的位置预测。请具体说明支持你地理结论的证据。" "city": "你预测的城市名称", "country": "你预测的国家名称", "continent": "你预测的大陆名称", "longitude": 预测的经度值, "latitude": 预测的纬度值 } 重要要求：

- 1。你必须回复一个完整的 JSON 对象——JSON 之外不得有任何解释，不得拒答，也不得声明不确定
- 2。如果不确定具体坐标，请尽可能给出最准确的数值估计
- 3。所有字符串值都应使用英文
- 4。经度应介于 -180 和 180 之间
- 5。纬度应介于 -90 和 90 之间
- 6。"reason" 字段必须包含对音频线索的详细逐步分析
- 7。即使音频听起来很普通，你仍必须给出一个位置猜测并说明推理过程

请听音频，并立即提供 JSON 回复：

B.5.2 贡献计算提示

贡献计算提示

你正在分析一段音频地理定位推理文本。你的任务是从所提供的列表中找出哪些声音类别对位置预测有所贡献，并评估其贡献程度（0.0 至 1.0）。可用的声音类别：{categories} 请分析推理文本，找出其中明确提及或暗示了哪些类别，以及每个类别对位置预测的贡献程度。请将回答输出为 JSON 对象，其中：- 每个键均为一个类别名称（必须与上方列出的名称完全一致）- 每个值均为 0.0 至 1.0 之间的贡献分数，其中：* 0.0 = 未提及/不相关 * 0.1-0.3 = 提及但贡献较小 * 0.4-0.6 = 贡献中等 * 0.7-0.9 = 贡献显著 * 1.0 = 主要/占主导的贡献仅包含贡献达到 > 0.0 的类别。只输出 JSON 对象，不要添加任何其他文本或说明。输出格式示例：{ "Speech": 0.8, "Music": 0.3, "Animal": 0.2 }

B.6 示例输出

在本节中，我们展示了所选模型针对三个代表性案例生成的完整输出。对于每段音频片段，我们还使用 NanoBanana Pro 生成了对应的图像，以便于解读和理解音频。

B.6.1 示例 1-Gemini 3Pro

示例1



{ “reason": "这段音频录音呈现出鲜明的声音特征，指向北非马格里布地区。最显著的初始线索是以该地区典型风格进行的宣礼（伊斯兰教宣礼），这立刻表明录音地点位于一个穆斯林占多数的国家。其背景中还叠加着海鸥持续而响亮的尖叫声，以及明显的风噪，这强烈暗示这里是一座濒临大型水域的沿海城市。交通噪声中，可以听到二冲程轻便摩托车或踏板车穿行于狭窄街道时特有的高音鸣响，这是摩洛哥麦地那常见的声景。录音的后半段，声音转为强劲而有节奏的海浪拍打海岸，听起来更像是大西洋汹涌的浪潮，而非地中海通常较为平静的海浪。从语言特征来看，可以听到“亚拉”（阿拉伯语）的呼喊，以及似乎是法语的短句，体现了摩洛哥的双语传统。强劲的风、随处可闻的海鸥、宣礼声以及汹涌的大西洋浪涛，这些特征共同构成了索维拉的标志性声景。索维拉位于摩洛哥，因其历史悠久的设防港口而闻名，并有“非洲风城”之称。"

"city": "索维拉", "
country": "摩洛哥", "
continent": "非洲", "
longitude": -9.7595, "
latitude": 31.5085 }

B.6.2 示例1-GPT-4o Audio

示例 1



{ “reason": "这段音频录音中清晰地传来了宣礼，也就是伊斯兰教宣礼，这种声音通常出现在穆斯林占多数的国家。背景噪音中有交通声和城市声响，表明这是城市环境。宣礼与城市噪音同时出现，因此可以推断，这可能是一个伊斯兰教占主导地位的国家中的城市。考虑到一些知名城市，土耳其的伊斯坦布尔是一个合理的猜测，因为它规模庞大，在穆斯林世界中具有重要地位，并融合了现代与传统。正如音频中所反映的那样，这座城市经常交通繁忙，而且其地理位置也与这些可听到的线索相符。" "city": "伊斯坦布尔", "country": "土耳其", "continent": "欧洲", "longitude": 28.9784, "latitude": 41.0082 }

B.6.3 示例 1-Qwen3-Omni

示例 1



{ “reason": "音频中混合着多种声音，强烈暗示某个特定的地理位置。最突出的声音是响亮、高音调且持续不断的汽笛声，这是日本列车或地铁广播的典型特征。随后传来车轮在

铁轨上有节奏的金属碰撞声，这是火车在铁轨上行驶时的典型声音。背景中还有一台大型强劲发动机的声音，可能来自渡轮或大型船只，同时还能听到海浪声。这些声音——列车／地铁、渡轮和海浪——结合起来，表明这里是一座沿海城市，且公共交通系统发达。背景中说的是日语。这些元素组合在一起，尤其是火车汽笛的独特声音和沿海环境，极具日本特色。日本神户是一座重要的港口城市，拥有发达的铁路网络，坐落于濑户内海沿岸，因此这里可能同时听到火车和渡轮的声音。"

"城市": "神户", "国家": "日本", "大陆": " 亚洲 ", " 经度 ": 135.198, "纬度": 34.693 }

B.6.4 示例 1-Mimo-Audio

示例 1



{ “reason": "音频中有多个独特线索，指向某个特定的地理位置。最主要的线索是录音中反复出现的火车汽笛声，说明录音可能是在铁路沿线录制的。此外，背景中还有交通声，包括汽车和可能的公共汽车，表明这是一个城市环境。环境噪声水平较高，混杂着机械声和人为声，这通常是繁华城市的特征。虽然没有清晰的人声或语言可以明确识别地点，但火车

声与城市交通噪声相结合，强烈暗示这里是一座铁路网络活跃的大城市。一个可能的地点是印度孟买，这座城市以密集的城市环境和发达的铁路系统而闻名。火车声和交通声都符合该市的基础设施特征。不过，由于缺少更具体的音频线索，这仍然只是一个有根据的猜测。"

"城市": "孟买", " 国家": "印度", "大陆": "亚洲", "经度": 72.8, "纬度": 19.0 }

B.6.5 示例 2-Gemini3 Pro

示例 2



{ "reason": "该音频环境是一个宽敞且回声明显的大厅，能听到行李箱滚动、车门关闭以及人群低语等独特声音，表明这里是一座繁忙的火车站。在0:48处，可以听到微弱的瑞典语对话。明确的地理证据始于1:08：先是一声标准的电子提示音，随后是一名女性通过自动广播用瑞典语播报车站广播。播报内容为：‘Till tåg [number]... mot Uppsala... via Arlanda C, Knivsta... Var god stig ombord, spår 6’（前往乌普萨拉的火车……途经阿兰达中央站、克尼夫斯塔……请登车，6号站台）。这条特定线路向北行驶，前往乌普萨拉市，并停靠阿兰达机场（阿兰达中央站）和克尼夫斯塔，始发于该地区的主要铁路枢纽。因此，这段音频录制于斯德哥尔摩中央车站，具体是在站台层。", "city": "斯德哥尔摩", " country": "瑞典", "continent": "欧洲", " longitude": 18.0583,

"latitude":
59.3301 }

B.6.6 示例 2-Qwen3-Omni

示例 2



{ "reason": "音频中包含多个关键线索，指向一个特定位置。最确凿的线索是用韩语播报的公告，而韩语是韩国使用的语言。公告内容与一列‘KTX’火车有关，这是韩国铁道公社在韩国运营的高速铁路服务。公告还提到了‘首尔站’，这是首尔的重要铁路枢纽。背景音包括火车在轨道上行驶时独特的轰鸣声和咔哒声、车门关闭的声音，以及火车站的整体环境音。综合韩语、对KTX和首尔站的提及，以及火车声，可以有力地表明这段录音是在韩国首尔站录制的。", "city": "首尔",

"country": "韩国", "
continent": "亚洲", "
longitude": 126.9781, "
latitude": 37.5326 }

B.6.7 示例 3-Gemini 3Pro

示例 3



{ "reason": "最突出的音频线索是乌鸫（Turdus merula）清晰、如笛声般悦耳动听的鸣唱

。这种鸟遍布欧洲，以适应城市环境的能力著称。录音的声学特征同样很能说明问题：鸟鸣和背景噪声都带有明显的混响或回声，这强烈暗示录音是在“内院”中录制的——这是一种由板式公寓楼围合而成的大型庭院。这种建筑风格是中欧城市的标志，柏林尤为典型。背景中持续传来遥远交通声形成的“城市嗡鸣”，没有刺耳的鸣笛声；开头的机械声（沉重的轰鸣和金属摩擦的尖啸）则像是附近驶过的有轨电车（Straßenbahn）或城市列车。特有的“庭院声学”、此起彼伏的乌鸫鸣唱，以及欧洲公共交通基础设施发出的声音，共同构成了鲜明的声音特征，广为人知，正是德国柏林住宅区的声景。",

"city": "柏林" "
country": "德国", "
continent": "欧洲", "
longitude": 13.405, "
latitude": 52.52 }

B.6.8 示例3-GPT-4o Audio

示例 3



{ "reason": "这段音频录音包含自然环境声音，其中尤为突出的是鸟叫声和啁啾声，听起来像是郊区或公园区域特有的声音。有时还能听到远处的交通噪声，表明附近有道路或城市环境。鸟的种类和整体声景表明当地气候温和，可能位于北半球。鉴于这些自然声

音没有具体语言或文化音频标志等鲜明特征，因此很难进行具体的位置预测。不过，考虑到温和的自然环境以及此类公园和环境的普遍性，我在考虑一座拥有广阔公园、气候温和的知名国际城市。因此，我猜测可能是纽约市，中央公园便是这类环境的典型地点。这个选择只是推测，依据是假设这里是一座拥有显著绿地、兼具城市与自然声音的大城市。", "city": "纽约市" "country": "美国", " continent": "北美洲", "longitude": -73.9654, "latitude": 40.7829 }

"country": "英国", " continent": "欧洲", " longitude": -0.1278, " latitude": 51.5074 }

B.6.9 示例 3-Qwen3-Omni

示例 3



{ "reason": "音频中有一段格外突出且复杂的鸟鸣，具有欧洲歌鸫（Turdus philomelos）的典型特征。这种鸟在欧洲很常见。背景主要是城市交通声，包括汽车声和清晰的车辆喇叭声。这些特定鸟鸣与交通噪声的组合强烈暗示地点位于西欧。整体声景是欧洲城市公园或绿地的典型环境。欧洲歌鸫的鸣唱与城市环境相结合，指向英国境内的某个地点；这种鸟在英国分布广泛，交通模式也与音频相符。伦敦是英国的一座大城市，拥有此类环境。", "city": "伦敦"