SCENEBench：面向辅助与工业应用场景的音频理解基准测试

LayaIyer[1] **Angelina Wang**[2] **Sanmi Koyejo**[1]1斯坦福大学 2康奈尔理工学院 laya, sanmi@stanford.edu angelina.wang@cornell.edu

arXiv:2603.09853v1 [cs.SD] 10 Mar 2026

摘要

大语言模型（LLM）的进展推动了音频处理能力的显著提升，催生了如今被称为大型音频语言模型（LALM）的先进模型。然而，除了自动语音识别（ASR）之外，衡量音频理解能力的研究仍然很少。本文提出基准测试套件 SCENEBench（空间、跨语言、环境及非语音评测），旨在填补这一空白，并针对四类真实场景中的广泛音频理解能力进行评测：背景声音理解、噪声定位、跨语言语音理解和声音特征识别。这四类任务的选取依据是无障碍技术和工业噪声监测中尚未得到充分研究的需求。除性能外，我们还测量模型延迟。该基准测试套件旨在评估音频中超越语词本身的内容——也就是说，不仅关注说了什么，还关注说话的方式以及音频中的非语音成分。由于我们的音频样本是合成构建的（例如，将两个自然音频样本叠加），我们还使用每项任务 20 个自然音频条目对基准进行进一步验证。这些条目从现有数据集中抽样，并符合我们的任务标准，以评估生态效度。我们评估了五种先进的大型音频语言模型，发现其中存在关键短板：各项任务的表现参差不齐，有些任务的性能低于随机猜测水平，另一些则具有较高的准确率。这些结果为有针对性地提升模型能力指明了方向。

1 引言

自然语言理解领域的显著进步推动了搜索、对话和信息检索等广泛应用的发展。随着这些模型能力的提升，我们必须开发有效评估它们的方法。

语音是人类广泛共享的交流方式，也是基于文本的大语言模型的延伸。然而，理解口语并不局限于转录。要实现恰当的音频理解，就必须识别语调、情绪、背景噪声、环境背景、说话者意图等。这些要素往往同时存在，并会显著影响含义。它们也使实用系统成为可能，例如辅助设备或字幕工具，可为听力障碍人士描述交通声音和接近的警笛；远程医疗和健康福祉系统则可检测患者语音中的咳嗽声或啜泣声。

发布大型音频语言模型的公司，例如 GPT-4o（OpenAI，2024年）和 Qwen2-Audio（Chu等人，2024年），宣传其具备超越自动语音识别的能力——例如，阿里巴巴声称，Qwen2-Audio“能够转录语音并识别音频信息……包括口语、音乐和环境噪声”（阿里巴巴云，2024年）。然而，目前的评估策略主要衡量语音识别能力，而非音频理解能力。也就是说，它们评估的是说了哪些词，而不是这些词是如何说出来的，也不关注音频中的非语音成分。

本文提出了一套新的基准测试套件 SCENEBench，用于评估大型音频语言模型更广义的音频理解能力，其设计依据是无障碍技术和工业噪声监测的需求。许多以往的基准测试侧重于受控、单模态或干净的场景。我们的基准测试涵盖四类反映现实世界复杂性的任务：背景声音理解、噪声定位、跨语言语音理解和人声特征描述（例如哭泣）。所有这些任务还会与模型延迟这一维度一并测量；模型延迟是我们基准测试套件中与其他维度并列的指标。

2 相关工作

在第 2.1 节中，我们回顾并梳理了以往音频与语音理解基准测试的范围（表 1）。在第 2.2 节中，我们通过两个影响重大且利害关系明确的应用场景——无障碍和工业监测——分析现有差距。我们还将探讨我们的基准测试套件如何推广到其他情境。

2.1 音频与语音理解基准测试

若干基准测试评估模型处理语音和音频输入的能力，各自采用不同的设计选择，以专注于其主要目标。

AudioBench（Wang 等人，2025）优先考虑广泛覆盖和可自动评分的任务。这种设计支持大规模、可复现的跨模型比较，并降低标注成本。然而，它弱化了那些较难捕捉和自动评分的现象。**MMAU**（Sakshi 等人，2024年）在各类任务中采用多项选择形式，以提高标注者间一致性、减少评估差异和提示敏感性，并简化对大型模型套件的评分。多项选择约束使结果具有可比性且稳定，但也限制了模型展示开放式推理、论证或描述细微声学属性的能力。**AIR-Bench**（Yang 等人，2024年）引入开放式音频问题，但重点更多放在音乐和环境声音上，而非对语音语用学和副语言学进行细致考察。**CAVA**（Held 等人，2025）针对语音助手行为（例如，指令遵循、延迟），契合实际部署中的关注重点。这一侧重点展现了智能体在约束条件下的表现和响应能力，但对细粒度声学理解的探查较少。其他工作，如 **SoundCheck**（Agnew 等人，2024年）和 **SONAR**（Li 等人，2025），则审计数据集或检测深度伪造，服务于特定目标，而非全面评估音频理解。

我们的基准测试套件旨在补充现有音频基准测试的覆盖范围，重点在于：为我们的四项任务增加有针对性的评估，并在需要时纳入自由作答评分。这些领域在现有基准测试的优先级设置中代表性不足。与以往基准测试的并列比较见表 1。这些缺口在高风险的现实部署场景中影响最大，我们将在下一节进一步讨论。

2.2 用例驱动的差距

从应用场景来看，现有音频基准测试仍偏重于“洁净室”语音识别或字幕生成场景，而很少涉及错误后果最为严重的两个领域：**(i) 无障碍**和**(ii) 工业声音监测**。

无障碍 日常场景中常有串扰、交通噪声、设备声，以及带有情绪或耳语的语音。对于聋人或听力障碍者（DHH）用户，有用的支持不应仅限于转录，还应具备声音感知能力（Jain 等人，2019；Bragg 等人，2016年）。需要凸显显著非语音事件的一些关键实例包括听辨警笛及其状态（接近还是远离），以及辨别副语言线索（疲劳、啜泣声、耳语）（Findlater 等人，2019；Wu 和 Jain，2025年；Kim 等人，2023年）。

已证明，能够检测警笛和发声并通过触觉或屏幕警报传递信息的可穿戴与物联网系统可以降低风险（Chin 等人，2023年；Salem 等人，2023年）。相反，背景噪声（串扰、宽带噪声）可能会降低自动语音识别的性能，而非使其更丰富，这凸显了模型理解情境背景而非仅识别前景词语的必要性（Dilmegani，2025年）。

工业监测 工厂车间和实验室需要尽早检测异常机器声音，而这些声音通常出现在语音或环境噪声背景下。ToyADMOS 和 MIMII 等公共数据集为机器状态监测中的声学异常检测（AAD）建立了通用测试平台（Koizumi et al.，2019；Purohit et al.，2019）。近期，MIMII-DG 引入了跨机器类型和录音设置的领域偏移，以测试这些模型的泛化能力，而这往往是部署中的关键阻碍（Dohi etal.，2022）。遗漏或误分类细微线索会直接转化为安全风险。

其他应用 远程医疗和身心健康领域也反复出现类似需求，例如咳嗽和呼吸道症状筛查、疲劳／嗜睡线索（Orlandic 等人，2021年；Bhattacharya 等人，2023年；Schuller 等人，2011年）；在智能家居／物联网领域，用于家庭活动和罕见事件检测，例如玻璃破碎声（Dekkers 等人，2018；Mesaros 等人，2017）；在交通运输和公共安全领域，例如城市场景中的警笛／喇叭声感知（Gemmeke 等人，2017）；以及增强现实／虚拟现实和机器人技术领域，在这些场景中，空间化音频线索支持导航和交互-

基准测试	Type	多turn	主要重点	音频片段 (≈)	#任务	Bkgd 声音	噪声 Loc.	跨语言	人声字符	延迟
SCENEBench	多项选择+自由作答	✓	音频与语音理解	16k	4	✓	✓	✓	✓	✓
音频Bench (Wang 等人，2025)	多项选择+自由作答	✗	自动语音识别、场景与语音	100k	8	✓	✗	✗	✗	✗
MMAU (Sakshi 等人，2024年)	MC	✗	多任务推理	10k	27	✓	✗	✗	✗	✗
AIR-Bench (Yang 等，2024年)	多项选择+自由作答	✓	生成式理解	21k	20	✗	✗	✗	✗	✗
MARBLE (Yuan 等，2023年)	MC	✗	音乐分类	25.9k	18	✗	✗	✗	✗	✗
Clotho-AQA (Lipping 等人，2022)	FRQ	✗	环境QA	1.9k	1	✓	✗	✗	✗	✗
SoundCheck (Agnew 等人，2024年)	审计	✗	数据集质量审计	3M	7	✓	✗	✗	✗	✗
SONAR (Li 等人，2025)	CLS	✗	深度伪造检测	2.2k	1	✗	✗	✗	✗	✗
CAVA （举行等人，2025）	Task	✓	语音助手行为	6.4k	6	✗	✗	✗	✗	✓
AU-Harness （Surapaneni等人，2025）	MC	✗	ASR, 场景 & 语音	5k	6	✓	✓	✗	✗	✗
LMMs-Eval (Zhang 等人，2025)	MC	✗	多模态评测	2k	8	✓	✗	✓	✗	✗

表 1：基准测试对比（采用标准化规模与覆盖范围）。类型：MC = 多项选择、FRQ = 自由作答、CLS = 分类、Audit = 数据集审计。音频片段（≈）表示已明确说明或可推导的条目总数；任务数统计不同的任务类别（未报告时以 — 表示）。覆盖范围列表示各套件是否评估相应维度。

化音频线索支持导航和交互（Chen 等人，2020，2021年）。

我们基准测试的设计依据。在这些设置中，四种失败模式反复出现：重要事件可能发生在背景中；音频理解会丢失空间／距离线索；遇到多语言片段时，转录质量会下降；非语音所承载的副语言信息缺乏明确界定。

SCENEBench 将每种失败模式转化为有针对性的评估。这四项任务不可或缺，因为它们检验了与我们的使用场景密切相关的关键音频理解能力，其设计也有具体的失败模式和潜在用户收益作为依据。本次发布暂不纳入完整的空间音频设置和长音频片段，以确保基准测试精简、可复现，并契合实时辅助和监控场景。总体而言，这些任务的选择构成了一套评估方案，能够诊断当前大型音频语言模型在哪些方面表现出色、又在哪些方面仍有不足。

选择这些任务的依据我们选择这四个评估维度并非随意为之。回顾前两节所述的文献时，相同的失败模式屡次出现：系统会忽略显著的背景声音（警笛、警报、机器音调），无法在噪声环境下追踪运动或距离线索，将多语言或语码转换的语音统一为单一语言，也会遗漏与情境感知相关的咳嗽声、笑声或耳语等非语音人声事件。SCENEBench 将这些已有文献记载的问题直接落实为四项任务——背景声音理解、噪声定位、跨语言语音理解和人声特征描述——以便系统地衡量当前大型音频语言模型在哪些方面表现不佳。我们认为，这套评测方案旨在诊断问题，而非穷尽所有可能：它聚焦于文献中最常被报告、影响最大的缺口，同时为未来扩展留出空间。

3 方法

我们的基准测试套件[1] 基于现有数据集构建，并对其应用特定任务转换。本节介绍四项任务各自的数据构建流程，并重点说明这些流程与源数据集的常规用途有何不同。我们评估了五种领先模型，包括 GPT-4o（OpenAI，2024年）、Qwen2-Audio-7B-Instruct（Chu等人，2024）和 Gemini-2.5（Comanici 等人，2025），以揭示当前

[1]数据和代码：https://github.com/layaiyer1/SCENEbench。

大型音频语言模型在音频优先场景中的局限性 .

3.1 任务

我们的基准测试包含四种任务类型，旨在探究真实场景下音频理解的不同维度。在 第 3.1.1 节中，我们通过在语音下方嵌入环境音频，并评估模型识别该音频的能力，来测试背景声音理解。 第 3.1.2 节介绍了一项基于振幅变化的 噪声定位任务，针对无障碍场景中的警笛检测等应用。 第 3.1.3 节通过要求模型转写多种语言的话语，考察其跨语言鲁棒性。最后，第 3.1.4 节聚焦于副语言线索，测试模型能否识别并标注人声特征描述，例如笑声和耳语。在 第 3.2 节中，我们介绍基准测试的一个并行维度——模型延迟，并在所有任务中对其进行测量，以评估实时性能约束。

3.1.1 背景声音理解

第一个任务考察模型能否识别叠加在语音下方的背景噪声。尽管先前的工作，如 WSJ0-2mix（Hershey 等，2016年；Isik 等，2016年）和 Libri2Mix（LibriMix）（Cosentino 等，2020）旨在进行说话人分离，但很少有基准测试探究语音下方的环境声音。因此，我们将 ESC-50 类别（Piczak，2015年）叠加到 DailyTalk 话语（Lee 等，2023年）上，并使用两种语音（较高与较低音高），为 2,000 个音频片段中的每个片段创建两个版本。背景噪声和语音均以原始音量叠加，其局限性将在第7节中讨论。评分采用分层方式：我们通过三个提示词评估背景噪声任务：（FR1）自由回答描述所有可听到的声音；（FR2）定向追问，明确询问背景声音（仅当 FR1 遗漏该信息时才发出）；以及（MC1）对背景类别进行四选一强制选择。评分采用分层方式：如果自由回答提及任意背景噪声类别，则 FR1（描述音频。）判定为正确；如果在 FR1 中（此时 FR2 无需提问即可获得满分）或 FR2 追问中说出了正确的 ESC-50类别，则 FR2（描述背景噪声。）计分；四选一多项选择测试针对所有音频片段进行，并独立评分。提示词的确切措辞见附录A。输出经过清理（转换为小写、去除标点），并与每个类别对应的同义词列表进行匹配，同时进行简单的否定/不确定性处理，例如“没有警笛声”“不是狗”“不确定”。这种三层设计有意将自发显著性（FR1）与定向检索（FR2）和可辨别性（MC1）区分开来，从而揭示失败究竟源于遗漏、误命名，还是在多个合理类别之间产生混淆。

3.1.2 噪声定位

该任务评估模型检测音频中动态音量模式的能力，通过振幅调制模拟空间运动。我们基于 ESC-50 环境声音语料库创建了一个数据集：对 2,000 个源音频片段应用三种不同的音量包络，共生成 6,000 个样本。每个原始声音均经历了三种变换：（1）接近中的声源，振幅在音频片段时长内从 10% 增加至 100%；（2）远离中的声源，振幅从 100% 线性缩减至 10%；（3）经过听者身旁的振荡运动，振幅在 20% 到 100% 之间按照正弦模式变化（完整循环 4 次）。

采用两种互补的提示词对模型进行评估。首先（FR1），通用描述提示词要求模型描述所有听觉特征。其次（FR2），后续位置提示词专门询问相对于声源的空间关系和运动模式。提示词的确切措辞见附录C。如果回答提及了恰当的运动模式或适当的同义词（例如“接近”“远离”“振荡”），则自动计为正确。

3.1.3 跨语言声音识别

我们通过将 DAILYTALK 转录文本转换为受控的混合语言刺激材料，来评估多语言片段转录：对于每个最长轮次（共 2,541 个），我们使用谷歌翻译将连续片段（≈30%）翻译成四种语言之一（普通话、西班牙语、印地语、葡萄牙语），仅保留通过回译相似度 > 0.9 检验的条目，并使用 ElevenLabs 多语言文本转语音合成音频。这些生成的音频文件包含多语言句子（例如：“我有十五天的假期，我想拥有一个去英格兰的旅行”），随后播放给各种大型音频语言模型进行转录评估（FR1），并通过模型转录文本与参考多语言句子之间的相似度来衡量性能。由于这些语言之间自然切换代码的高质量人类录音十分稀缺，我们

采用这种合成方案以实现覆盖面和可控性；我们在此简要说明其局限，并在 第7节 中进行讨论。因此，这项任务——多语言片段转录——是代码切换行为的一种代理指标。

3.1.4 人声特征描述

我们关注无需进行情感推断、却能传递交流线索的非言语发声特征——咳嗽、哭泣、笑声、打喷嚏、打哈欠、含糊低语和耳语。鉴于已有研究记录了简化标签可能引发的伦理问题及其潜在危害（Stark 和 Hoey，2021），我们有意避免直接进行情绪分类。相反，我们的评估要求模型先简要描述每个音频片段（FR1），再对这些人声类别进行七分类（MC1）。该数据集整合了公开可用的资源：NON- SPEECH7K 包含咳嗽/哭泣/笑声/打喷嚏/打哈欠（W4ng1204，2023），CAPSPEECH-AGENTDB- AUDIO 包含含糊低语/耳语（OpenSound，2024）；此外，还从 VO- CAL BURSTS TAXONOMY 100 CLEAN WDS _ _ _ _ _ 中补充了含糊低语（Krishnakalyan，2023），并从 ASMR 中补充了耳语（Nyuuzyou，2022）。最终数据集包含 4,006 个音频片段，涵盖报告的五种标签（632 个咳嗽、1,791 个哭泣、1,133 个笑声、236 个打喷嚏、214 个打哈欠），另有含糊低语和耳语用于七分类。

3.2 延迟这一维度

除了准确率外，我们还报告了本地模型的延迟；云端/API模型不纳入计时比较。对于每次模型调用（一次提示词—音频查询），我们记录从发起调用到测试框架收到完整文本响应之间的时长。我们使用中位数和四分位距汇总所有四项任务的延迟，并报告各任务的中位数，以展示延迟如何随输入内容和提示词类型而变化。

3.3 评测模型

我们对五种具备音频能力的先进大型音频语言模型进行基准测试：GPT-4o（OpenAI，美国）（OpenAI，2024年），Gemini 1.5（Google DeepMind，美国/英国）（Comanici 等人，2025），Qwen2-Audio（阿里巴巴达摩院，中国）（Chu等人，2024年），Audio-Flamingo-3（NVIDIA，美国）（Goel 等人，2025），以及 DeSTA2-8B-beta（国立台湾大学 + NVIDIA，台湾）（Lu 等人，2024年）。这些模型涵盖了多种架构、训练范式和地理来源。

我们选择这五种模型，以覆盖与我们的任务最为相关的设计空间，并在公平性与可复现性之间取得平衡：它们涵盖商业模型（GPT-4o、Gemini）和开放权重模型（Audio-Flamingo-3、Qwen2-Audio-7B、DeSTA2-8B-beta）；而且它们是近期被广泛使用的基线模型，声称具备与我们的任务相符的多语言和非语音能力。在本文中，我们排除了仅支持语音的自动语音识别模型和音乐专用模型，因为它们无法在不引入额外组件的情况下运行完整测试套件，而这些组件会干扰比较结果。

4 结果

我们报告了四种任务类型的结果（摘要见 图1），每种任务旨在探究音频理解的一个不同维度。在 第4.1节中，我们分析模型检测叠加在语音之下的背景声音的能力。 第4.2节评估模型估计背景噪声方向的能力。 第4.3节介绍多语言语境下转写准确率的研究结果。最后，第4.4节 考察模型对非语音发声（如笑声和咳嗽声）的识别能力，以评估其对情感及副语言声音的理解。

4.1 背景声音理解

每个音频片段都包含前景语音和 ESC–50 背景声音。模型首先给出自由回答（FR1：“描述该音频。”）。如果 FR1 未提及背景声音，我们会提出指定的后续问题（FR2：“说出任意一种具体的背景声音。”）。最后，模型回答一个四选一的多项选择题（MC1）。评分时，如果模型在 FR1 中已经回答正确，即使没有提问 FR2，我们也将 FR2 计为正确（因此所有音频片段都会计入 FR2）。“不确定/无法判断”标记为错误。柱状图显示平均值及 95%置信区间；水平虚线表示随机猜测水平（MC1 为 25%）。为便于阅读，此处报告关键结果；完整的统计细节见附录B。95%置信区间极窄（$\leq \sim$1在$N=4000$处半宽为 0.5 个百分点），因此在柱状图上无法直观区分。为完整起见，相关数据收录于附录B。

延迟。 对于有计时日志的模型，我们统计 FR1 + FR2（仅当 FR1 失败时） + MC1。Flamingo 速度很快（中位数**2.26**秒；第90百分位数为 2.73 秒），而 Desta 较慢（中位数**15.61**秒）。GPT–4o 和 Gemini 未纳入统计，因为它们是基于 API 的模型。

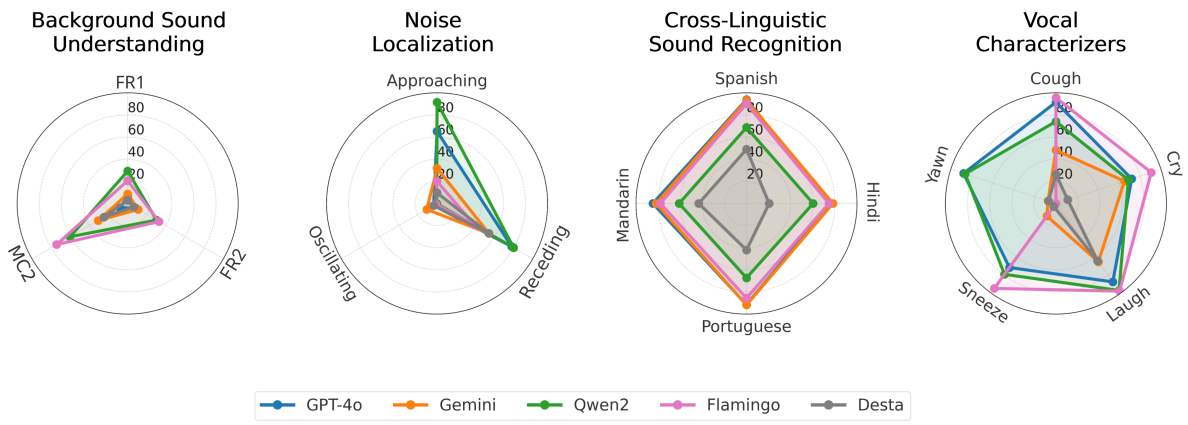

图1：各任务的汇总雷达图。 每个轴代表一个任务特定类别；数值为百分比（或平均相似度百分比）。图例仅显示一次：GPT-4o（蓝色）、Gemini（橙色）、Qwen2（绿色）、Flamingo（粉色）、Desta（灰色）。

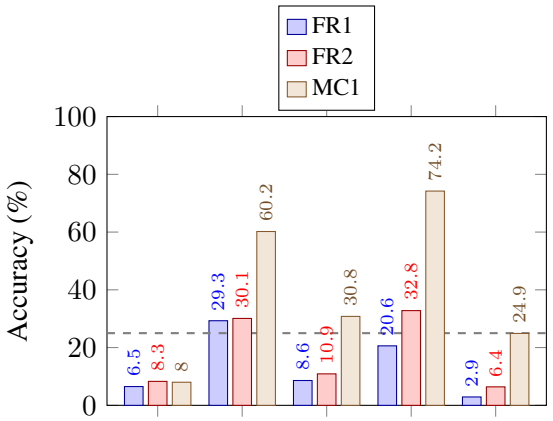

图2：各模型的环境声准确率 （N=4000 个音频片段）。“特定类型*”将 FR1 正确数视为 FR2 正确数。简单随机的基线为 25%（多项选择；4 选 1）。

模型很少会主动提及背景噪声（FR1），但在使用显式提示（FR2）或提供选项（MC1）时，表现会好得多。Flamingo 在 FR2/MC1 上表现最佳；通义千问 Qwen2 在 FR1 上表现最佳。提出直接、具体的问题可显著提升表现。

4.2 噪声定位

我们通过对 ESC-50 音频片段施加振幅包络（接近 ↑amp、远离 ↓amp、振荡正弦波），测试模型能否检测运动。每个音频片段对应两种提示词：FR1（总体）自由文本描述；FR2（方向）明确询问位置/运动。若回答指出真实运动，即视为正确。

FR1：总体描述 所有模型在 FR1 上仍面临挑战。按类别来看，远离始终是 FR1 中最容易识别的类别（例如 Gemini 为 27.4%），而振荡对所有模型而言都接近最低水平（≤10%）。

模型	正确数	准确率（%）	95%置信区间
Gemini	981	16.35	[15.4, 17.3]
GPT-4o	658	10.97	[10.20, 11.78]
Flamingo	481	8.02	[7.49, 8.58]
Desta	362	6.03	[5.51, 6.65]
Qwen	447	7.45	[6.81, 8.14]

表 2：运动 FR1（总体描述） — 准确率及 Wilson 95%置信区间； $N = 6000$ 各模型。

FR2：方向当我们启用“FR1 已正确时， ⇒ 在 FR2 上自动纠正 ” 规则，并明确询问方向时，表现最好的模型得分大幅提升：

模型	正确数	准确率（%）	95%置信区间
Qwen	3434	57.23	[55.98, 58.48]
GPT-4o	2938	48.97	[47.70, 50.23]
Gemini	1934	32.23	[31.06, 33.43]
Desta	1344	22.40	[21.36, 23.47]
Flamingo	1209	20.15	[19.11, 21.23]

表3：运动FR2（方向）——准确率及 Wilson 95%置信区间； $N = 6000$ 每个模型。根据设计，FR1正确的计数也计为FR2正确。

直接询问（FR2）有帮助： Δ（FR2–FR1）=+38.0分（GPT-4o）， +49.8分（通义千问），+15.9（Gemini）， +12.1（Flamingo）， +16.4（Desta）。模型难以在自由文本（FR1）中自发给出运动线索，但在被明确询问时（FR2）往往能够作答。振荡仍是一个尚未解决的情形。

4.2.1 延迟

我们报告每个片段的“有效”延迟 =FR1延迟 + 若FR1失败则计入FR2延迟（若未询问FR2则为0）。记录的（本地）中位数：Flamingo 2.32秒；通义千问6.04秒；Desta 14.53秒。本次运行未对API模型（GPT-4o、Gemini）计时。完整延迟表见附录D。

4.3 跨语言评估

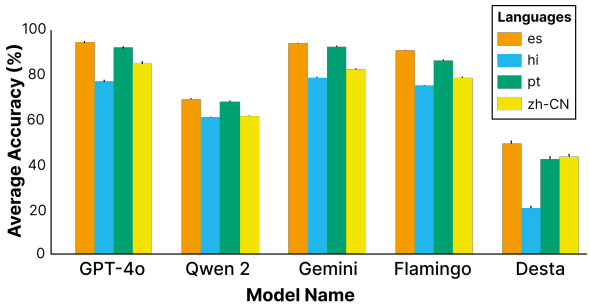

图 3：跨语言转录准确率（平均值± 95% CI），按模型和语言划分。 语言：西班牙语 (es)、印地语 (hi)、葡萄牙语 (pt)、普通话 (zh-CN)。各语言的片段数量为 N_{es}=1010, N_{hi}=1034 N_{pt}=1052 $N_{zh\text{-}CN}$=884 ,,。误差棒为基于每段片段相似度得分的正态近似置信区间。

我们衡量转录 相似度（0–1；以百分比报告），采用多语言 DailyTalk 音频片段（构建方法见第3.1.3 节）。各语言平均值的 95% 置信区间非常窄（通常为 <1 个百分点），因此在图 3 的柱状图中可能难以辨认。为完整起见，每个模型在各语言上的确切置信区间范围见 × 。附录F。

GPT–4o 和 Gemini 在不同语言上轮流领先（西班牙语/葡萄牙语成绩接近；印地语中 Gemini 领先 →；普通话中 GPT–4o 领先 →）。Flamingo 在西班牙语/葡萄牙语上的表现略逊于领先者；Qwen2 处于中游；Desta 表现最低。所有模型在普通话上的表现都未达到上限。

仅本地（开放权重）模型：Flamingo 中位数 **0.92** 秒（第90百分位数 1.23 秒），Qwen2 **1.41** 秒（第90百分位数 1.92 秒），Desta **4.93** 秒（第90百分位数 10.13 秒）。本次运行未计时 API模型（GPT–4o、Gemini）。

4.4 人声特征描述

非言语发声（咳嗽、哭泣、笑声、打喷嚏、打哈欠）用于测试对声学形态的识别能力，而不依赖语言内容。我们在**4,006** 个**五选**一音频片段上进行评估（632 个咳嗽、1,791 个哭泣、1,133 个笑声、236 个打喷嚏、214 个打哈欠）。为使表格简洁，我们报告**平均值** ±**95% CI**。作为简单随机参考，五选一的基线为 20%。

模型	咳嗽	Cry	笑声	打喷嚏	Yawn
GPT-4o	91.5 ±2.2	72.1 ±2.1	87.6 ±1.9	71.6 ±5.8	87.9 ±4.4
Gemini	48.1 ±3.9	65.2 ±2.2	65.3 ±2.8	14.0 ±4.5	7.5 ±3.6
Qwen2	74.1 ±3.4	69.0 ±2.2	97.2 ±1.0	79.2 ±5.2	86.4 ±4.6
Flamingo	95.6 ±1.6	90.7 ±1.4	98.0 ±0.8	94.9 ±2.9	0.0 ±0.9
Desta	26.3 ±3.4	11.2 ±1.5	64.1 ±2.8	3.4 ±2.4	7.0 ±3.5

表 4：发声特征识别准确率（百分比）。数值为平均值 ± 95% 置信区间。

有效延迟为描述时间与多项选择时间之和 +（仅限本地模型）：Flamingo **0.80**秒（第90百分位数 0.97），Qwen2 **1.22**秒（第90百分位数 1.74），Desta **6.84**秒（第90百分位数 62.52）。本次测试未对 API模型（GPT–4o、Gemini）计时。Flamingo 在咳嗽、哭泣、笑声和打喷嚏任务上表现领先，但无法识别打哈欠；GPT–4o 和通义千问在全部五类任务上表现出色；Gemini 表现不一，在打喷嚏和打哈欠任务上的表现低于随机猜测水平；Desta 则落后。所有模型的总体表现均高于 20% 的随机猜测水平。

4.5 自然样本

为评估生态效度，我们还对一个规模较小的自然产生的人类录音数据集进行了评估[2]，包含每项任务 N=20 个音频片段，其标签和问题格式与合成基准相同（构建详情见 H 节）。这些样本并非通过合成修改音频而来，而是从现有数据集中选取，以满足各项任务所需的声学特征。因此，人类片段在背景噪声、说话人位置和语言混用方面呈现出更自然的变化。

在分析了Salaudeen 和 Hardt (2025) 对排名稳定性的研究后，我们考察了模型在合成数据与人工标注数据上的相对排序是否保持一致。总体而言，虽然各项任务的准确率绝对范围与合成结果有所不同，但相对性能模式通常保持不变，如 图 4所示：在合成数据上表现出色的模型，在人类录音上往往仍然表现强劲，而较弱的模型则依然表现不佳。

在各项任务中，人类录制音频片段的绝对表现与合成基准测试有所不同，但模型的整体相对排名基本保持不变（图 4），这表明合成评估能够反映有意义的比较趋势。

[2]与主数据集一同发布。

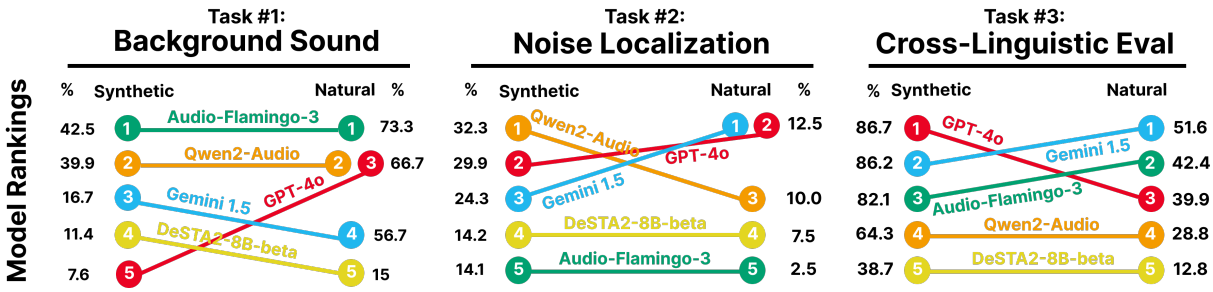

图 4：合成数据和自然数据上的模型排名。 尽管各任务的模型排名在合成评估和人工录制评估中有所不同，但在合成数据上表现出色的模型，在自然录音上的平均表现通常也依然强劲。

模型在自然音频的背景声音理解方面表现更好，这可能是因为显著事件是间歇性出现的，而非完全混杂在一起；而噪声定位表现下降，则是由于空间线索不那么明显，且语音的存在会使模型倾向于转录，而非进行空间推理。人类录音中的跨语言准确率同样有所下降，这反映出更大的语言变异性和有限的语码转换；尽管包含嵌入的英语内容，模型仍会固守主导语言。我们没有为人声特征描述单独设置人工验证集，因为数据集中的这些音频片段直接取自真实录音，而非合成生成。

每项人工验证任务的详细信息和数据见第I节。

4.6 错误分析

错误类型	背景。声音	噪声 Loc.	跨语言	人声字符。
遗漏	18%	43%	–	1%
过于笼统的标签	4%	–	–	–
错误归因	3%	21%	–	94%
方向互换	–	8%	–	–
省略语言	–	–	64%	–
部分再翻译	–	–	36%	–
噪声覆盖	75%	–	–	–

表5：各任务的错误类型分布（占汇总数据集中错误案例的百分比）。所列百分比是以该任务中的错误为条件计算的，而非以所有音频片段为基数。标记为“–”的单元格表示该错误类别不适用于该任务。

除了总体准确率之外，我们还进行了结构化错误分析，以了解模型在各项任务中的失误方式。针对每项任务，我们从不同模型的错误案例中抽取了分层随机样本（每个模型、每项任务抽取 10 个音频片段；共 $N=200$ 个）。随后我们基于这组样本建立了错误分类体系，并按照该体系为全部样本标注。我们的标签包括：(i)遗漏（从未提及目标事件），(ii)过度泛化标签（例如，用“交通”代替具体的 ESC-50类别），(iii)错误归因（归为错误的具体类别），(iv)方向互换（接近↔远离），(v)省略语言（省略非英语片段），(vi)部分再翻译（只翻译外语片段），以及 (vii)噪声覆盖（只关注音频片段中的语音内容，而忽略背景线索）。错误类型并非互斥。

背景声音的主要失败类型是噪声覆盖：在 75% 的错误自由回答中，模型转录了语音，却完全没有提及背景事件。单纯遗漏占错误的 18%，过度泛化标签（例如 “ 背景噪声 ”“ 交通”）占 4%，错误归因占 3%。

在噪声定位任务中，主要问题是遗漏（43%；未描述运动）、振荡塌缩（25%；周期性的音量强弱模式被简化为“音量变化”）以及运动类型误归因（21%）。方向互换（接近↔远离）占 8%，短时程聆听占 2%。

在跨语言转录中，64% 的错误属于单语规范化，即非英语片段被删去并替换为流利的英语；36% 属于部分再翻译，即尽管指令要求保留语码混用，却只翻译了外语片段。

在人声特征描述方面，94% 的错误是非语音类别之间的错误归因（例如，打哈欠与叹息），提及我们五个标签之外的人声特征描述的情况占 5%，遗漏约占 1%。

综合来看，这些模式表明模型偏重前景语音而忽视背景事件，对运动的时间推理能力有限，并且会对

多语言片段进行规范化处理，同时对非语音人声线索的标注不稳定。

5 讨论

SCENEBench 通过评估大型音频语言模型（LALM）能否在真实环境中检测显著的背景事件、推理运动、保留多语言片段，并识别非语音发声，来探究其音频能力。我们的结果喜忧参半。

在各项任务中，模型在大多数多项选择格式上的表现接近或超过随机猜测水平，但通常会遗漏辅助和工业应用场景中真正重要的现象。在背景声音理解方面，遗漏占据主导：模型几乎从不在自由文本中主动提及背景事件，即使通过针对性问题或选项提示，也只能部分恢复。在噪声定位方面，当我们明确询问运动情况时，模型的表现显著提升，但仍会将振荡包络混为一谈，并过度关注音频片段的结尾。在跨语言转录方面，最常见的错误是将非英语片段完全规范化掉。在人声特征描述任务中，即使非词汇事件（例如打哈欠）才是关注的标签，词汇内容仍会压过副语言线索。

这些模式表明，当前的大型音频语言模型针对说了什么（自动语音识别和字幕生成）进行了优化，而非如何表达或场景中还发生了什么。这并不令人意外：大多数公开的音频基准测试都聚焦于纯净语音、单一声源的环境音频片段或音乐分类，并围绕易于自动评分的任务设计。因此，模型即使在现有测试套件中得分很高，仍可能无法掌握音频理解的基本要素。

那么，为什么这些缺口一直被忽视？我们的分析表明，这些能力之所以受到的关注较少，存在若干结构性原因。首先，收集并标注重叠事件（带有背景音频、运动和副语言线索的语音）比整理干净的单标签音频片段困难得多；因此，大多数广泛使用的语料库根本不包含这类内容。此外，词错误率、描述文本BLEU分数或单标签准确率等常见指标会奖励词汇层面的忠实度和泛化的场景标签，却不会因模型遗漏警笛、将运动合并，或消除多语言片段而对其进行惩罚。SCENEBench 正是围绕这些“棘手”案例设计的，并非为了取代现有评测套件，而是为了让人们更难在不处理这些问题的情况下，声称具备全面的音频理解能力。

提升这些任务表现的一种方法，是聚焦于有针对性的数据和训练目标。对于背景声音理解，一种改进现有模型的方法是让模型接触带有背景声音描述的语音与噪声混合音频。具体而言，我们可以加入以下信息：（i）是否存在背景类别；（ii）为其命名；同时提供困难负样本，使前景叙述正确、背景标签错误，因为大型音频语言模型在自动语音识别方面表现良好（Chu等人，2025；Wang 等人，2025；Yang 等人，2024年）。对于定位任务，使用物体运动已知的音频片段进行训练；加入用于分类接近／远离／振荡的目标，并确保模型能够随时间整合证据，而不是过度看重最后几秒。对于多语言片段，在片段标注转录文本上进行指令微调和偏好微调，并使用对比样本对，其中唯一的区别是是否保留或翻译外语片段。对于副语言学，进行微调或引入包含嵌入语音中的简短非词汇事件的数据。在各项任务中，采用多任务微调，将这些目标与标准的自动语音识别／字幕生成结合起来，确保词汇忠实度，同时让模型学会关注音频特征。

为跟踪这些针对性改动在实践中是否确实有所帮助，我们的基准测试提供了相应的评估信号。总之，该套件能够识别当前模型擅长和不足之处，为引导训练和模型设计从解析语音转向解析音频提供了简洁、可复现的测试平台，而不仅仅是处理语音。

5.1 未来工作

为扩大覆盖范围，未来工作可优先考虑以下三项改进：

- **自然语码转换数据**。 将我们第三项任务中的合成生成数据替换为一个涵盖多种语言对的小型人工录制语料库，以验证（或修正）从合成数据集中得出的结论，并更好地反映真实的语码转换行为。
- **真实声学环境**。 超越等响度叠加与合成运动，具体做法是：(i) 扫描语音与背景信噪比（例如，-10 至 $+10$dB）；(ii) 加入含移动声源（警笛/车辆/机械设备）的现场录音，以测试接-

近/远离在多普勒、遮挡及移动听者情形下的表现。

- **更强的基线模型**。纳入非大型音频语言模型（LALM）流程以便对结果进行对照（例如，语音分离 →ESC–50 分类器，用于背景声音理解）；这为鲁棒性比较明确划定了“传统方法能达到什么水平”的基准线，也让我们能够指出，结果反映的是任务难度还是模型局限。

6 结论

总体而言，**SCENEBench** 揭示了具体的失败模式——遗漏、过度泛化标签、归因错误、短时程聆听以及语言先验主导——这些问题在更干净的单事件基准测试中往往被掩盖。我们提供了适合各项任务的随机基线，报告了由大规模 N带来的紧密置信区间，并将完整统计数据列于 附录 B。我们希望这些评估能引导模型开发，使其朝着无障碍和工业安全所需的能力迈进。

7 局限性

我们的研究存在若干局限性。首先，混合设计依赖于受控电平的语音-噪声叠加。尽管等电平混合便于实验控制，但它可能无法反映现实环境中遇到的信噪比分布。其次，部分上游语料库包含较弱或不完善的标注。残留的标签错误可能影响性能上限估计，并增加各类别之间的混淆。第三，闭源模型 API 限制了细粒度的延迟分析和消融研究。因此，我们尽可能报告计时延迟指标。最后，多语种语音设置采用了带回译过滤的文本转语音（TTS）流程，这提高了一致性，但未能充分体现真实人类语音的自然随意性。

8 伦理考量

我们还强调了若干伦理考量。为降低简化式“情绪 AI”带来的风险，我们避免直接进行情感推断，而是聚焦于副语言事件。所有语音数据（包括合成混音）都必须遵守许可限制，任何人类录音都必须获得明确同意，并遵循数据最小化原则。必须谨慎考虑无障碍方面的风险：在警笛检测等安全关键场景中，漏检和误报都可能造成不同的伤害。因此，系统应公开经过校准的不确定性，并提供回退行为。最后，公平性至关重要。基准测试应扩大对不同口音、语言和录音条件的覆盖范围，以减少不同用户群体之间的差异化错误率。

披露：大语言模型被用于润色正文和表格。

SK部分得到美国国家科学基金会2046795和2205329、美国教育科学研究所R305C240046、ARPA-H、麦克阿瑟基金会、Schmidt Sciences、斯坦福HAI、RAISE Health、OpenAI、微软和谷歌的支持。

参考文献

William Agnew、Julia Barnett、Annie Chu、Rachel Hong、Michael Feffer、Robin Netzorg、Harry H. Jiang、Ezra Awumey和Sauvik Das。2024年。SoundCheck：审计音频数据集。*arXiv*预印本 *arXiv:2410.13114*。

阿里云。2024年。阿里云发布Qwen2-Audio模型以分析语音和音频。访问日期：2025年10月6日。

Debarpan Bhattacharya、Neeraj Kumar Sharma、Debot- tam Dutta、Srikanth Raj Chetupalli、Pravin Mote、Sri- ram Ganapathy、C. Chandrakiran、Sahiti Nori、K. K. Suhail、Sadhana Gonuguntla 和 Murali Alagesan。2023年。Coswara：用于 SARS-CoV-2感染远程筛查的呼吸音与症状数据集。*Scientific Data*，10(1):397。Danielle Bragg、Naomi Huynh 和 Richard E. Lad- ner。2016年。面向聋人及听障用户的可个性化移动声音检测应用设计。收录于第*18*届 *ACM* 无障碍计算国际会议（*ASSETS*）论文集，第3–13页。Changan Chen、Ziad Al-Halah 和 Kristen Grauman。2021年。语义音视导航。预印本，arXiv:2012.11583。Changan Chen、Unnat Jain、Carl Schissler、Sebastia Vi- cenc Amengual Gari、Ziad Al-Halah、Vamsi Krishna Ithapu、Philip Robinson 和 Kristen Grauman。2020。Soundspaces：3D环境中的音视导航。预印本，arXiv:1912.11474。Chiun-Li Chin、Chia-Chun Lin、Jing-Wen Wang、Wei- Cheng Chin、Yu-Hsiang Chen、Sheng-Wen Chen 和 Po-Cheng Hsieh。2023年。采用边缘计算识别警笛的听力障碍者可穿戴辅助设备。《*Sensors*》，23(17):7454。Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、Yichong Leng、Yuanjun Lv、Jinzheng

He、Junyang Lin、ChangZhou 和JingrenZhou。2024年。Qwen2-Audio技术报告。预印本，arXiv:2407.10759。

Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen、Luke Marris、Sam Petulla、Colin Gaffney、Asaf Aharoni、Nathan Lintz、Tiago Cardal Pais、Henrik Jacobs- son、Idan Szpektor、Nan-Jiang Jiang及另外 3416 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。预印本，arXiv:2507.06261。

Joris Cosentino、Manuel Pariente、Samuele Cornell、Antoine Deleforge和Emmanuel Vincent。2020年。LibriMix：一个开源数据集，旨在实现通用语音分离。*arXiv*预印本 *arXiv:2005.11262*。包含 Libri2Mix 和 Libri3Mix 变体。

Gert Dekkers、Lode Vuegen、Toon van Waterschoot、Bart Vanrumste 和 Peter Karsmakers。2018年。DCASE 2018 挑战赛——任务 5：基于多通道声学的家庭活动监测。预印本，arXiv:1807.11246。

Cem Dilmegani。2025。2025年语音识别的四大挑战与解决方案。
h
t
t
p
s://
research.aimultiple.com/speech-recognition-challenges/。访问日期：2025-07-24。Kota Dohi、Tomoya Nishida、Harsh Purohit、Ryo Tanabe、Takashi Endo、Masaaki Yamamoto、Yuki Nikaido和Yohei Kawaguchi。2022。MIMII DG：面向领域泛化任务的故障工业机器调查与检测声音数据集。 预印本，arXiv:2205.13879。Leah Findlater、Bonnie Chinh、Dhruv Jain、Jon E. Froehlich、Raja Kushalnagar和Angela Carey Lin。2019。 聋人和听力障碍人士对可穿戴及移动声音感知技术的偏好。载于*2019*年*CHI*计算系统人因会议论文集（*CHI '19*），第1–13页，美国纽约州纽约。ACM。Jort F. Gemmeke、Daniel P. W. Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R. Channing Moore、Manoj Plakal和Marvin Ritter。2017。Audio Set：面向音频事件的本体与人工标注数据集。载于*2017 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第776–780页。Arushi Goel、Sreyan Ghosh、Jaehyeon Kim、Sonal Kumar、Zhifeng Kong、Sang gil Lee、Chao-Han Huck Yang、Ramani Duraiswami、Dinesh Manocha、Rafael Valle和Bryan Catanzaro。2025。Audio Flamingo 3：借助完全开放的大型音频语言模型推进音频智能。 预印本，arXiv:2507.08128。Will Held、Michael J. Ryan、Aditya Shrivastava、Ali Sartaz Khan、Caleb Ziems、Ella Li、Martijn Bartelds、Michael Sun、Tan Li、Woody Gan和Diyi Yang。2025。 CAVA：语音助手综合评估。
https://github.com/SALT-NLP/CAVA。一项在六个领域评估大型音频模型（LAMs）能力的基准测试：轮流发言、指令遵循、函数调用、语调感知、安全性和延迟。

John R. Hershey、Zhuo Chen、Jonathan Le Roux 和 Shinji Watanabe。2016年。深度聚类：用于分割和分离的判别式嵌入。载于神经信息处理系统进展（*NeurIPS*）。介绍了用于单通道语音分离的 WSJ0-2mix 评估设置。

Yusuf Isik、Jonathan Le Roux、Zhuo Chen、ShinjiWatan- abe 和 John R. Hershey。2016年。使用深度聚类进行单通道多说话人分离。载于*INTERSPEECH* 会议论文集。

Dhruv Jain、Leah Findlater 和 Jon E. Froehlich。2019年。探索聋人或听力障碍者在家中的声音感知。载于 *2019*年 *CHI*计算机系统人为因素会议论文集，第 94:1–94:13 页。

JooYeong Kim、Sooyeon Ahn和Jin-Hyuk Hong。2023年。 可见的细微差异：一种面向聋人及听障人士、用于可视化副语言语音线索的字幕系统。收录于 *2023*年*CHI*计算机系统人为因素会议论文集（*CHI '23*），美国纽约州纽约。ACM。Yuma Koizumi、Shoichiro Saito、Hisashi Uematsu、Noboru Harada和Keisuke Imoto。2019。ToyADMOS：用于异常声音检测的微型机器运行声音数据集。收录于 *IEEE WASPAA*会议论文集。Kalyan Krishnakalyan。2023年。人声爆发分类法100干净版（wds）。
h
t
tps://huggingface.co/datasets/krishnakalyan3/vocal_bursts_taxonomy_100_clean_wds。访问日期：2025年9月6日。Soket Labs。2025。 Coshe-eval：面向印地语-英语语音的语码转换自动语音识别基准。Hugging Face数据集。Keon Lee、Kyumin Park和Daeyoung Kim。2023年。DailyTalk：用于对话式文本转语音的口语对话数据集。收录于 *IEEE ICASSP*会议论文集。Xiang Li、Pin-Yu Chen和Wenqi Wei。2025。 SONAR：合成AI音频检测框架与基准测试。Samuel Lipping、Parthasaarathy Sudarsanam、Konstantinos Drossos和Tuomas Virtanen。2022。 Clotho-AQA：用于音频问答的众包数据集。 *arXiv*预印本*arXiv:2204.09634*。Ke-Han Lu、Zhehuai Chen、Szu-Wei Fu、Chao-Han Huck Yang、Jagadeesh Balam、Boris Ginsburg、Yu-Chiang Frank Wang和Hung-yi Lee。2024年。在没有语音指令微调数据的情况下开发指令遵循语音语言模型。 *arXiv*预印本 *arXiv:2409.20007*。

Dau-Cheng Lyu、Tien-Ping Tan、Eng Siong Chng和 Haizhou Li。2010。 SEAME：东南亚的普通话-英语语码转换语音语料库。收录于*Interspeech 2010*，第1986–1989页。

Annamaria Mesaros、Toni Heittola 和 Emmanouil Benetos。2017年。DCASE 2017 挑战赛任务 2：在真实场景音频中检测罕见声音事件。载于*DCASE* 研讨会论文集。

Nyuuzyou。2022年。ASMR 耳语音频片段合集。https://huggingface.co/datasets/nyuuzyou/asmr。Hugging Face数据集；访问日期：2025年10月6日。

OpenAI。2024年。GPT-4o 系统卡。https://cdn.openai.com/gpt-4o-system-card.pdf。访问日期：2025年7月24日。

OpenSound。2024年。CAPSPEECH-AGENTDB-AUDIO：含糊低语和耳语音频片段。https://huggingface.co/datasets/OpenSound/CapSpeech-AgentDB-Audio。Hugging Face数据集；访问日期：2025年10月6日。

Lara Orlandic、Tomislav Teijeiro 和 David Atienza。2021 年。CoughVid 众包数据集：用于 COVID-19 咳嗽分类的语料库。*Scientific Data*，8(1)：156。Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur。2015 年。LibriSpeech：基于公有领域有声书的自动语音识别语料库。载于 *2015* 年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 5206–5210 页。IEEE。Karol J. Piczak。2015 年。 ESC-50：环境声音分类数据集。载于 *ACM* 多媒体会议论文集，第 1015–1018 页。Harsh Purohit、Ryo Tanabe、Kohei Ichige、Tomoya Nishida、Takashi Endo、Yuma Koizumi、Noboru Harada、Masahiro Yasuda 和 Satoru Tamura。2019。MIMII 数据集：用于故障工业机器调查与检测的声音数据集。载于*DCASE* 研讨会论文集。arXiv：1909.09347。S. Sakshi、Utkarsh Tyagi、Sonal Kumar、Ashish Seth、Ramaneswaran Selvakumar、Oriol Nieto、Ramani Duraiswami、Sreyan Ghosh 和 Dinesh Manocha。2024 年。 MMAU：大规模多任务音频理解与推理基准测试。 *arXiv* 预印本 *arXiv*：*2410.19168*。Olawale Salaudeen 和 Moritz Hardt。2025 年。 ImageNot：与 ImageNet 形成对比可保留模型排名。预印本，arXiv：2404.02112。Osman Salem、Ahmed Mehaoua 和 Raouf Boutaba。2023 年。 助听视觉系统：一种帮助听力障碍驾驶员的基于物联网的系统。载于 *2023* 年 *IEEE* 计算机与通信研讨会（*ISCC*）。IEEE。访问地址：https://rboutaba.cs.uwaterloo.ca/Papers/Conferences/2023/SalemIoT2023.pdf。

Björn Schuller、Stefan Steidl、Anton Batliner、Florian Schiel 和 Jarek Krajewski。2011 年。Interspeech 2011计算副语言学挑战赛：醉态、困倦和年龄。载于*INTERSPEECH* 论文集。

SDAIA。2022年。沙特语代码切换语料库（SCC）。https://www.kaggle.com/datasets/sdaiancai/saudilang-code-switch-corpus-scc。知识共享署名-非商业性使用-相同方式共享 4.0 国际许可协议。

Luke Stark 和 Jevin D. Hoey。2021年。人工智能系统中的情绪伦理。*ACM* 人机交互会议论文集，5(CSCW2)：1–34。讨论了简化式情绪识别的风险及其相关伤害。

Sidharth Surapaneni、Hoang Nguyen、Jash Mehta、Aman Tiwari、Oluwanifemi Bamgbose、Akshay Kalkunte、Sai Rajeswar 和 Sathwik Tejaswi Madhusudhan。2025。Au-Harness：用于整体评估音频大语言模型的开源工具包。预印本，arXiv:2509.08031。

W4ng1204。2023年。Nonspeech7k：非语音发声数据集（咳嗽、哭泣、笑声、打喷嚏、打哈欠）。https://huggingface.co/datasets/W4ng1204/Nonspeech7k。Hugging Face数据集；访问日期：2025年10月6日。Bin Wang、Xunlong Zou、Geyu Lin、Shuo Sun、Zhuohan Liu、Wenyu Zhang、Zhengyuan Liu、AiTi Aw 和 Nancy F. Chen。2025年。AudioBench：音频大语言模型通用基准测试。*arXiv* 预印本 *arXiv:2406.16020*。2025年5月最后修订。Liang-Yuan Wu 和 Dhruv Jain。2025。Soundnarratives：支持聋人及听力障碍者的丰富听觉场景描述。载于第*27*届 *ACM SIGACCESS* 计算机与无障碍国际会议论文集（*ASSETS '25*），美国纽约州纽约市。ACM。发表于 ASSETS 2025；预印本可获取。Qian Yang、Jin Xu、Wenrui Liu、Yunfei Chu、Ziyue Jiang、Xiaohuan Zhou、Yichong Leng、Yuanjun Lv、Zhou Zhao、Chang Zhou 和 Jingren Zhou。2024年。AIR-Bench：通过生成式理解评测大型音频语言模型。*arXiv* 预印本 *arXiv:2402.07729*。获 ACL 2024 接收。Ruibin Yuan、Yinghao Ma、Yizhi Li、Ge Zhang、Xingran Chen、Hanzhi Yin、Le Zhuo、Yiqi Liu、Jiawen Huang、Zeyue Tian、Binyue Deng、Ningzhi Wang、Chenghua Lin、Emmanouil Benetos、Anton Ragni、Norbert Gyenge、Roger Dannenberg、Wenhu Chen、Gus Xia 及其他6位作者。2023年。MARBLE：用于通用评估的音乐音频表征基准。*arXiv* 预印本 *arXiv:2306.10548*。NeurIPS 2023 的定稿版本。Kaichen Zhang、Bo Li、Peiyuan Zhang、Fanyi Pu、Joshua Adrian Cahyono、Kairui Hu、Shuai Liu、Yuan-han Zhang、Jingkang Yang、Chunyuan Li 和 Ziwei Liu。2025。LMMs-Eval：对大型多模态模型评估的现实检验。载于论文集：*Findings of the*

计算语言学协会：*NAACL 2025*，第881–916页。计算语言学协会。

A 背景声音任务提示词

FR1（自由回答描述）。

> 请描述你在此音频文件中听到的内容。请尽可能详细地说明你能识别出的所有声音。

FR2（定向追问；仅在 FR1 遗漏背景/环境声音时提出）。

> 在你之前的回答中，你没有提到背景噪声。请具体描述你在此音频中听到的背景噪声或环境声音。

多项选择（四选一多项选择；对所有音频片段实施）。

> 你将收到一段音频片段。背景声音是什么？请从以下选项中选择正确答案，并且只回复数字（例如 1、2、3 或 4）。

B 背景声音任务的统计方法

我们使用威尔逊区间计算各模型比例的 95%置信区间。FR1（任何噪声）— GPT–4o：6.5% [5.73，7.27], 通义千问Qwen2：29.3% [27.89，30.71], Gemini：8.6% [7.73，9.47], Flamingo：20.6% [19.35，21.85], Desta：2.9% [2.38，3.42]。FR2（特定类型*）— GPT–4o：8.3% [7.44，9.16], 通义千问Qwen2：30.1% [28.68，31.52], Gemini：10.9% [9.93，11.87], Flamingo：32.8% [31.35，34.25],Desta：6.4% [5.64，7.16]。多项选择 — GPT–4o：8.0% [7.16，8.84], 通义千问Qwen2：60.2% [58.68，61.72],Gemini：30.8% [29.37，32.23], Flamingo：74.2%[72.84，75.56], Desta：24.9% [23.56，26.24]。

针对每个模型和多项选择题（MC1），我们使用单样本比例 z 检验来检验 H_0: $p = p_{\text{chance}}$，其基准值为 0.25（MC1；1/4）。为控制 15 项检验（5 个模型 $\times$ 3 指标）中的多重比较，我们采用 Benjamini–Hochberg 错误发现率方法，设定为 α=0.05。效应量以绝对风险差 $\Delta = p - p_{\text{chance}}$ 和 Cohen's $h = 2\arcsin\sqrt{p} - 2\arcsin\sqrt{p_{\text{chance}}}$ 报告。样本量和检验效能说明：对于 N=4000 个音频片段，当检验效能为 0.8 时，MC1 的最小可检测 $|\Delta|$ 为 <2 个百分点。

多项选择（MC；基线 25%）。Flamingo**74.2%**，通义千问2 60.2%，Gemini 30.8%，Desta 24.9%，GPT–4o 8.0%。Flamingo、通义千问2 和 Gemini**高于随机猜测水平**（BH校正 $p<10^{-10}$）；Desta**与随机猜测水平无显著差异**（BH校正$p\approx0.83$）；GPT–4o**低于随机猜测水平**（BH校正$p<10^{-10}$）。

C 噪声定位任务提示词

FR1（自由回答描述）。

> 请描述您在此音频文件中听到的内容。请尽可能详细地描述您能辨认出的所有声音，包括音量或空间特征方面的任何变化。

FR2（定向追问；仅在 FR1 未说明噪声方向时提问）。

> 在这段音频中，我们相对于声源的位置如何变化？请描述空间关系以及你检测到的任何运动模式。

D 噪声定位统计与延迟

统计检验。对于每个模型和问题，我们检验 H_0: $p = \frac{1}{3}$ (33.3%)，使用单样本比例 z检验； p值在10项检验（5个模型 $\times$ 2 个问题）中采用 Benjamini–Hochberg 方法进行校正。所有模型在 FR1 上的表现都远低于随机猜测水平，或在 FR2 上远高于随机猜测水平，且所有基线检验均显著（BH 校正后$p < 10^{-4}$）。随后，我们进行所有模型的两两比较（双比例检验，在每个问题内进行 BH 校正），并报告完整结果：在*FR1*（一般场景）中，总体准确率较低（6–16%），但几乎所有模型对之间的差异都具有统计显著性；唯一 不显著的比较是 Flamingo3 与 Qwen2（$\Delta = +0.57$个百分点, z=1.16, p_{BH}=0.245），其余差异均为 $\Delta \in [1.4, , 9.7]$ 个百分点，且 p_{BH} <0.01；差距最大的是 Gemini 与 DESTA2（$\Delta\approx10.32$个百分点, $z=-17.92$, $p_{\text{BH}} < 10^{-15}$）。在*FR2*（位置）中，每组模型之间的差异都很大且显著；例如，Qwen2 与 Flamingo3（$\Delta\approx41.2$个百分点, $z=-47.06$, $p_{\text{BH}} \approx 0$）、Qwen2 与 GPT-4o（$\Delta\approx10.72$个百分点， z=11.74, $p_{\text{BH}} \approx 0$），甚至差距最小的 DESTA2 与 Flamingo3（$\Delta\approx3.32$个百分点， z=4.85, $p_{\text{BH}} \approx 1.24\times10^{-6}$）之间的差异也仍然显著。补充材料中报告了效应量，分别为绝对差值 $\Delta = p - \frac{1}{3}$ 和 Cohen's h。

各类别说明。方向准确率主要集中在接近/远离；振荡接近最低水平（例如， ≤10.8%）。

延迟表。

模型（本地）	中位数（秒）	平均值（秒）	第90百分位数（秒）
Flamingo	2.32	2.35	2.98
Qwen	6.04	6.54	10.25
Desta	14.53	27.96	72.16
GPT-4o	n/a	n/a	n/a
Gemini	n/a	n/a	n/a

表6：动作任务有效延迟（本地推理）。数值为每段片段中位数、平均值和第90百分位数，单位为秒；API模型（GPT-4o、Gemini）在本次运行中不可用。

E 跨语言任务提示

FR1（自由回答描述）。

> 请转录以下音频，保留其中的任何语码混用或多语言内容。如果音频同时包含英语和其他语言，请在转录文本中保留语码混用风格。请逐字输出转录文本，包括任何非英语单词或短语。

F 跨语言详情

统计。 我们通过正态近似计算片段级相似度的 95%置信区间；对于每种语言的模型间比较，我们使用双样本 t 检验对每段片段得分进行比较，并在模型对之间进行 Benjamini–Hochberg 校正。效应量以均值差（百分点）及其 95%置信区间报告。

语言（N）	模型	平均值 ± 95% CI (%)
西班牙语（es, $N = 1010$）	GPT–4o	**93.9 [93.4, 94.4]**
	Gemini	93.3 [92.9, 93.7]
	Flamingo	90.1 [89.5, 90.7]
	Qwen2	68.7 [68.0, 69.4]
	Desta	49.0 [47.4, 50.5]
印地语 (hi, $N = 1034$)	Gemini	**78.2 [77.5, 78.8]**
	GPT–4o	76.5 [75.6, 77.5]
	Flamingo	74.6 [74.0, 75.1]
	Qwen2	60.3 [59.8, 60.9]
	Desta	20.5 [19.1, 21.8]
葡萄牙语 (pt, $N = 1052$)	Gemini	**91.8 [91.4, 92.3]**
	GPT–4o	91.5 [91.0, 92.0]
	Flamingo	85.6 [84.7, 86.4]
	Qwen2	67.3 [66.6, 68.1]
	Desta	42.1 [40.6, 43.6]
普通话 (zh–CN, $N = 884$)	GPT–4o	**84.7 [83.9, 85.4]**
	Gemini	81.8 [81.2, 82.4]
	Flamingo	78.0 [77.3, 78.7]
	Qwen2	61.0 [60.2, 61.7]
	Desta	43.2 [41.8、44.7]

表 7：跨语言转录（平均相似度 ±95%置信区间，%）。展示了每种语言对应的所有模型；每种语言的最佳平均值以粗体标出。

G VocalCharacterizers 任务提示

MC1。

> 以下哪一项最能描述这种声音？(A) 咳嗽 (B) 哭泣 (C) 笑声 (D) 打喷嚏 (E) 打哈欠 (F) 含糊低语 (G) 耳语。请用字母和词语作答。

H 人工标注样本数据集构建

为了评估模型在自然录音条件下的表现，我们从公开可用的真实世界音频数据集中选取音频片段，构建一个小型人工标注评估子集。针对每项任务，我们使采样和预处理流程与相应的合成基准测试保持一致，以确保可比性。

背景声音片段取自 Au- dioSet（Gemmeke 等人，2017年），并使用其中经过整理的声音事件标签。我们选取包含明确标注的背景事件的音频片段，并将每段录音裁剪为以标注事件为中心的 5 秒片段。此过程可生成前景语音与环境声音的自然混合，而无需叠加合成音频。

噪声定位样本取自 ESC50（Piczak，2015年）和 LibriSpeech（Panayotov 等人，2015年）。我们找出声音强度呈单调上升、下降或振荡变化的片段。将录音裁剪为能够捕捉这些振幅动态的短片段，以近似合成基准测试所用的运动线索，同时保留真实的声学变化。

多语种语音样本采自三个公开的语码转换语料库。普通话—英语片段取自 SEAME 语料库（Lyu 等人，2010），使用 HuggingFace 划分`AudioLLMs/seame_dev_sge`。沙特阿拉伯语—英语样本来自 SCC 数据集（SDAIA，2022），其中包含自发的沙特阿拉伯语语音，并频繁夹杂英语语码转换。英语—印地语片段取自 CoSHE-Eval数据集（Labs，2025），该数据集专为双语及语码转换语音识别而设计。

此外，每个音频片段均由不同的人说话，因此自然呈现出音高、口音和说话风格的差异，从而提升生态效度。

I 自然录音表现

背景声音理解 在人工录制的音频片段上，模型的表现优于合成音频片段。在 FR1 中，平均准确率从合成音频上的 13.6% 提升至人类

录音上的 45.0%，而 FR2 则从 17.7% 提升至 50.0%。MC1 的准确率也从 39.6% 提升至 72.0%。这些提升部分得益于人类片段的结构：在许多样本中，背景声音与语音同时出现，但并未覆盖整个录音。这种时间上的分离可能使背景事件更加明显，让模型比在完全混合的合成片段中更容易检测到它们。尽管绝对表现大幅提升，相对表现模式总体上仍保持一致：表现出色的合成模型依然表现良好，而 Desta 则始终表现不佳。

Task	模型	准确率（%）	95%置信区间
FR1	GPT-4o	60.0	[38, 79]
FR1	Gemini	45.0	[25, 66]
FR1	Flamingo	70.0	[47, 87]
FR1	Desta	0.0	[0, 17]
FR1	Qwen	50.0	[29, 71]
FR2	GPT-4o	70.0	[47, 87]
FR2	Gemini	45.0	[25, 66]
FR2	Flamingo	70.0	[47, 87]
FR2	Desta	5.0	[1, 24]
FR2	Qwen	60.0	[38, 79]
MC1	GPT-4o	70.0	[47, 87]
MC1	Gemini	80.0	[58, 92]
MC1	Flamingo	80.0	[58, 92]
MC1	Desta	40.0	[22, 61]
MC1	Qwen	90.0	[68, 98]

表8：任务1：背景声音理解 — 准确率及威尔逊95%置信区间；$N = 20$ 每个模型。

噪声定位模型在 FR1 的人工标注样本上表现不佳，平均准确率仅为 3%，相比之下，在合成数据上的准确率为 9.8%。在 FR2 中明确询问时，模型在真人录音上的准确率升至 15.0%，但仍低于合成数据 36.2% 的平均值。与合成数据集不同，合成数据通过受控的振幅调制刻意夸大运动线索，而真实录音中响度和空间位置的变化更加细微。此外，许多音频片段包含语音，而非像合成数据那样只有环境噪声。音频语言模型似乎优先处理转录，而非空间推理，因此对相对运动线索的关注度较低。

Task	模型	准确率（%）	95%置信区间
FR1	GPT-4o	5.0	[1, 24]
FR1	Gemini	5.0	[1, 24]
FR1	Flamingo	0.0	[0, 17]
FR1	Desta	0.0	[0, 17]
FR1	Qwen	5.0	[1, 24]
FR2	GPT-4o	20.0	[8, 42]
FR2	Gemini	20.0	[8, 42]
FR2	Flamingo	5.0	[1, 24]
FR2	Desta	15.0	[5, 36]
FR2	Qwen	15.0	[5, 36]

表9：任务2：噪声定位— 准确率及 Wilson 95%置信区间；$N = 20$ 每个模型。

跨语言评估在人类样本中，句子级相似度低于合成数据。按语言对取平均后，表现最佳的模型达到中等准确率：Gemini 为 51.6%，Flamingo 为 42.4%，GPT-4o 为 40.0%。通义千问的平均准确率为 28.8%，而 Desta 的表现明显较差，为 12.8%。这些差异既反映了更高的语言多样性，也反映了自然语音中语码转换较少：人类片段通常是在较长的非英语片段中嵌入较短的英语片段。因此，模型更可能继续使用主导语言，而无法识别嵌入的英语内容。这种行为与错误分析部分中观察到的“语言丢失”错误类别一致，第4.6节，不过在自然录音中出现得更频繁。尽管准确率整体下降，各模型的表现排序仍保持不变。

语言对	模型	准确率（%）	95%置信区间
英译中	GPT-4o	57.5	[35.3, 74.9]
英译中	Gemini	44.6	[31.8, 56.7]
英译中	Flamingo	63.6	[44.8, 80.7]
英译中	Desta	26.8	[10.0, 45.4]
英译中	Qwen	44.7	[32.1, 55.7]
英语-南非荷兰语	GPT-4o	25.6	[7.8, 48.6]
英语-南非荷兰语	Gemini	61.4	[44.7, 78.3]
英语-南非荷兰语	Flamingo	23.7	[14.4, 33.4]
英语-南非荷兰语	Desta	9.8	[2.9, 19.0]
英语-南非荷兰语	Qwen	17.3	[8.6, 27.1]
英语-印地语	GPT-4o	36.8	[20.0, 52.5]
英语-印地语	Gemini	48.8	[40.5, 56.5]
英语-印地语	Flamingo	40.0	[23.8, 53.8]
英语-印地语	Desta	1.7	[1.2, 2.1]
英语-印地语	Qwen	24.3	[14.8, 35.5]

表10：任务3：跨语言评估 —人类录音上的平均相似度准确率，置信区间为 95 $N =$ %（每个模型）。