MusicListenBench：音频大语言模型终于能听出差异了吗?

AbhinabaRoy[1], **Ambuj Mehrish**[2] **& Dorien Herremans**[1]1新加坡科技设计大学 2威尼斯大学
{abhinaba roy,dorien herremans}@sutd.edu.sg,ambuj.mehrish@unive.it

arXiv:2611.29375v1 [cs.SD] 22 Nov 2026

摘要

音频大型语言模型（LLM）通常依据歌曲流派、情绪和乐器编配进行评测。一项更基础的能力，是判断两个短音频片段播放的是不是同一段旋律、同一个和弦或同一种乐器。我们推出*MusicListenBench*，这是一个针对旋律、和声、音色和节奏的配对听力测试基准。每个测试条目都有两个变体。翻转变体会改变音乐，因此答案也必须改变。保持变体只改变声音，例如房间回声或嘶声，因此答案必须保持不变。这两个变体的答案相反，所以忽略音频的模型无论偏好哪个选项，得分都是 50%。该基准包含一个训练集划分和两个测试集：一个由符号音乐渲染而成，带有精确标签；另一个则由真实录音构成。在渲染测试中，八个开源音频大语言模型中有七个的得分仅比这一随机水平高出不到 4 个百分点。四个商业模型中，两个 GPT 音频模型都处于随机水平；表现最好的 Gemini 2.5 Pro 得分为 74.2%，比人类听众（87.0%）低 12.8 个百分点。使用 1 万对渲染的干净音频进行 GRPO 后训练，使 Qwen2.5-Omni 在留出干净配对上的得分从 61.2% 提升到 99.0%，而且这一提升也迁移到了训练中未涉及的任务。然而，在翻转/保持测试中，它的表现仅达到我们听众的水平（86.0%）。它几乎不再漏判声音变化（1.9%，而 Gemini 2.5 Pro 为 24.7%），但当声音变化而内容不变时，它有 25.7% 的概率把正确答案改成错误答案。其错误翻转中，超过一半是由轻微的背景嘶声引起的。模型学会听出差异的速度，远快于学会判断哪些差异重要。在真实录音上，开源模型的表现再次接近随机水平；经训练的模型则将 42 对由同一和弦扫弦两次构成的音频中的 27 对判为不同。将仅声音配对加入训练，可降低两个测试集上的错误翻转，但无论怎样训练，都无法教会模型辨别相对音高。我们发布了数据[1],代码[2], 以及一个开放排行榜[3]。

1 引言

音频大语言模型（音频大语言模型）可以描述录音的流派、情绪和乐器等特征（Yang 等人，2024年；Sakshi 等人，2025年；Weck 等人，2024年）。这些描述背后有一项更基础的能力：判断两个声音的某些特征是否相同。两段短音频片段演奏的是
同一旋律吗？是同一和弦吗？是同一种乐器吗？哪一段速度更快？这类相同/不同问题是听觉研究中的标准工具（Hautus 等人，2021年），人类音乐能力测试也由此构建（Law 与 Zentner，2012年；Wallentin 等人，2010年）。人类听众答对的比例为 87.5%。如果模型无法回答这些问题，就难以弄清它对音乐的描述依据是什么。MusicListenBench 针对的是这一层次的聆听能力，即音乐耳测试和 PROMS 所测的能力，而不是对整首歌曲的分析，例如判断其调、曲式或流派。这一层次是基础：如果模型无法判断两个短动机是否相同，就不能相信它能够描述一首完整歌曲的旋律。

近期的基准测试发现，音频大语言模型在基本的音乐与声音感知任务上处于或接近随机水平，例如判断旋律的音高是否发生了偏移（Carone 等人，2026年；2025年；Sun 等人，2026年）。目前仍存在两个问题。首先，相同/不同评分混合了两种能力：注意到重要的变化，以及

[1]https://huggingface.co/datasets/amaai-lab/MusicListenBench[2]
https://github.com/AMAAI-Lab/MusicListenBench[3]
https://amaai-lab.github.io/MusicListenBench/

忽略不重要的变化。这类评分还会受到文本先验以及偏好选择某个答案字母的影响而被夸大（Weck 等人，2024年；Zang 等人，2025年；L´opez 等人，2025年；Zheng 等人，2024年）。其次，这些基准测试只评估模型当前的状态，因此无法判断训练能否解决这一问题。

我们介绍 MusicListenBench（图 1）。它基于一个简单的想法：能够听出不同之处的模型必须做到两点：音乐发生变化时改变答案，而只有声音发生变化时保持答案不变。因此，每个测试条目都有两个变体。在翻转变体中，我们改变查询属性，例如旋律中的一个音符，正确答案也随之切换。在保持变体中，我们只改变音乐的听感，例如加入房间回声或噪声，而正确答案保持不变。这两种变体提出相同的问题，但答案相反。因此，忽略音频的模型无论偏好哪个选项，得分都是 50%（在三个相同/不同任务中恰好如此，节奏任务的平均得分也是如此；见第 3.3 节）。保持不变的配对还杜绝了一种干净配对所允许的捷径。对于干净的相同/不同配对，模型可以只检查两个音频片段听起来是否相似来作答，而无需知道旋律或和弦是什么。在保持不变的配对中，即使答案是“相同”，两个音频片段听起来也不同。主要测试集由符号音乐渲染而成，因此每个答案都确切已知；其中使用的音高范围、音色库和音效在训练中从未出现过。第二个测试集则使用真实录音构建相同类型的条目。

我们分四步使用该基准测试来回答标题中的问题。(1) 还不能。 在翻转/保持任务中，8 个开放音频大语言模型里有 7 个，以及两个 GPT-Audio 模型，准确率都仅比 50% 的随机水平高或低不到 4 个百分点。表现最好的商业模型 Gemini 2.5 Pro（74.2%）仍比我们的听众（87.0%；第5.1节）低 12.8 个百分点。(2) 训练可提升干净准确率。在我们的训练集划分上进行 GRPO，使 Qwen2.5-Omni 在留出干净配对上的准确率从 61.2% 提升至 99.0%。训练还提升了一项未纳入训练的任务（和声：从 59.2% 提升至 96.4%），同时使字幕生成表现接近基底模型（第5.2节和第5.4节）。(3) 模型学会察觉变化的程度，远大于学会忽略变化的程度。 在翻转/保持任务中，同一模型的准确率也仅达到我们的听众水平（86.0%）。它的漏检率从 36.7% 降至 1.9%，但错误翻转率仅从 41.0% 降至 26.1%，并未低于未经训练的 Gemini 2.5 Pro（26.9%）；而且几乎所有错误翻转都发生在模型对干净音频能正确作答的题目上。模型学会忽略训练中从未见过的均衡器变化（错误翻转从 33.7% 降至 1.4%），但面对背景嘶声或移调时则做不到。在音色和节奏任务中，我们的听众每项最多只错一道题，而嘶声导致了模型 94 次错误翻转中的 85 次。监督微调（SFT）也呈现出相同的模式。错误翻转与冻结的音频编码器有关：在渲染片段中，轻微的嘶声对其输出造成的影响，甚至大于音乐本身的真实变化。加入我们训练集划分中的仅声音配对后，错误翻转有所减少，即使是训练中从未见过的变化也如此（第5.3节）。(4)同样的问题也出现在真实音频中。 在真实测试集上，开放模型的表现再次接近随机水平，而使用我们的合成配对进行训练仍有帮助（从 52.1% 提升至 62.8%）。训练后的模型能在真实的扫弦和弦中检测出 41 个移位三度音中的 38 个，但也将 42 组由同一和弦的两次扫弦构成的配对中的 27 组判为不同。

我们的贡献包括：(i) MusicListenBench：包含 10,000 个训练项目和两个测试集，其中有 3,000 个带有精确标签、难度分级、针对捷径的检查以及留出测试条件的渲染项目，以及 1,800 个来自真实录音的项目；同时发布生成器、评分脚本和公开排行榜（第 3.5 节）；(ii) 翻转/保持协议，用于区分注意到重要变化与忽略无关变化；(iii) 将八个开源音频大语言模型和四个商业音频大语言模型与人类听众进行评估比较；(iv) 针对四个模型开展训练研究，采用 GRPO 和监督微调，考察向未见任务的迁移，以及对字幕生成、冻结的音频编码器、仅声音配对训练的副作用，并在 Slakh2100 上重复实验（附录 J）。

2 相关工作

音频大语言模型的基准测试。AIR-Bench、MMAU 和 MMAR 通过数千个有关语音、声音和音乐的问题测试音频大语言模型（Yang 等人，2024年；Sakshi 等人，2025年；Ma 等人，2026年），而 MuChoMusic 则专注于音乐（Weck 等人，2024年）。这些基准测试使用真实音频，以及由人类或大语言模型编写的问题，因此覆盖面广，但文本先验和字母偏差也可能影响得分（Weck 等人，2024年；Rouditchenko 等人，2025年；Zang 等人，2025年；L´opez 等人，2025年）。PitchBench（Dujardin 等人，2026年）通过 28 项对照实验测试音高听辨，这些实验改变响度、音符时长、声源、时间拉伸和背景噪声。CMI-Bench（Ma 等人，2025年）将音乐信息检索（MIR）数据集的标注转化为针对真实录音的指令任务，并使用标准音乐信息检索指标评分。还有三个基准测试与我们的工作最为接近。MUSE 测试十种音乐感知能力

其中包括相同/不同判断，使用 200 段录制的刺激音频和 200 名人类听众进行测试（Carone 等人，2026年）。Carone 等人（2025年）比较 MIDI 和音频输入在切分音、移调和和弦性质方面的表现。SonicBench 为十二种物理属性渲染受控声音，包括音色和节奏，并像我们一样将两个音频片段放在同一个文件中（Sun 等人，2026年）。这三种基准测试都只评估模型的原始表现，不进行训练。我们在三个方面有所不同。每个条目都有一个翻转变体和一个保持变体，因此我们可以区分模型是注意到了变化，还是忽略了变化；忽略音频的模型得分为 50%。全部 3,000 个渲染测试条目都有精确标签和分级难度，另一个来自真实录音的测试集则采用相同的构建方式。此外，我们还提供一个包含留出测试条件的训练集划分，以便基准测试衡量后训练带来的变化。Music Arena（Kim 等人，2026年）等实时竞技平台通过人类投票为文本生成音乐系统排名，评估的是模型生成的内容，而非它听到的内容。

配对测试。在听觉研究中，相同/不同判断和二选一设计是标准做法（Hautus 等人，2021年）。音乐耳测试会询问听众：两段短旋律是否相同，即使其中一个音符可能发生了变化（Wallentin 等人，2010）；这与我们的旋律任务很相似，而 PROMS 则采用旋律、节奏和音色的相同/不同判断试次（Law 与 Zentner，2012年）。针对模型，MelodySim 和 MERIT 会学习判断两个音频片段的旋律或其他音乐方面是否相似（Lu 等人，2025年；Roy 等人，2026a）；而我们让音频大语言模型用文字作出这一判断。在语言和视觉领域，最小对（Warstadt 等人，2020年）、对比集（Gardner 等人，2020年）以及配对图像描述测试（Thrush 等人，2022年）用于检验模型的答案是否会随输入中的正确变化而改变。MMVP 和 NaturalBench 会对需要不同答案的图像进行配对，只有当一对或一组中的所有答案都正确时才给分（Tong 等人，2024年；Li 等人，2024年）。CheckList 将测试分为两类：方向性测试中，编辑应当改变输出；不变性测试中，编辑则不应改变输出。翻转和保持是这两类测试在听觉任务中的对应形式。将它们配对还能抵消语言模型在选择题中表现出的选项字母偏差（Zheng 等人，2024年）。

用于音频和音乐模型的强化学习。Herremans 与 Roy（2026年）回顾了音乐生成中的偏好对齐及其尚待解决的问题，其中包括：当音乐质量具有主观性时，偏好模型的可靠性如何；以及推理时对齐如何提升符号音乐生成（Roy 等人，2026年b）。这类工作优化的是模型创作的内容；我们关注的是模型能否聆听——这是任何此类评判器首先需要具备的技能。组相对策略优化（GRPO；Shao 等人，2024年）已提升音频问答能力，例如 Qwen2-Audio 在 MMAU 上的表现（Li 等人，2025年）。

3 MusicListenBench

3.1 题目与任务

每个题目由一个音频文件和一个问题组成。文件中包含片段 A、1.0 秒的静音和片段 B。像 SonicBench（Sun 等人，2026年）那样将两个音频片段放在同一个文件中，使我们能够测试仅接受单个音频输入的模型。经过 GRPO 训练后，有一个模型能以这种格式正确回答 99.0% 的干净题目（第 5.2 节），因此这种格式本身并不会妨碍模型比较音频片段。每项任务的问题有 8 到 10 种人工撰写的表述。问题末尾有一条固定指令，说明哪个字母对应哪个答案，模型则以一个字母作答。表 1 列出了四项任务。旋律、和声和音色任务询问两个音频片段是否相同。节奏任务询问哪个音频片段更快（有些表述询问哪个更慢）。旋律、和声和节奏任务各有多个难度等级，测试题目均匀分布在这些等级中。

3.2 精确标签与无明显线索

我们使用 music21 将每个条目写成符号音乐（Cuthbert 和 Ariza，2010），并使用 FluidSynth（FluidSynth 开发者，2000）和通用 MIDI 音源库将其渲染出来。对于每个生成项目，我们会渲染一小组文件：基底配对、一个会改变答案的编辑后的片段，以及两个或三个只改变声音的音频片段。由于音频是生成的，每个答案都能被精确确定，无需人工标注。我们还会移除容易识别的线索。任务中的所有音频片段时长都完全相同，因此节奏更快的音频片段包含更多点击声，而不是文件更短。所有音频片段的响度都相同（−23响度单位）。按照标准的相同/不同设计，干净音频对中的两个音频片段

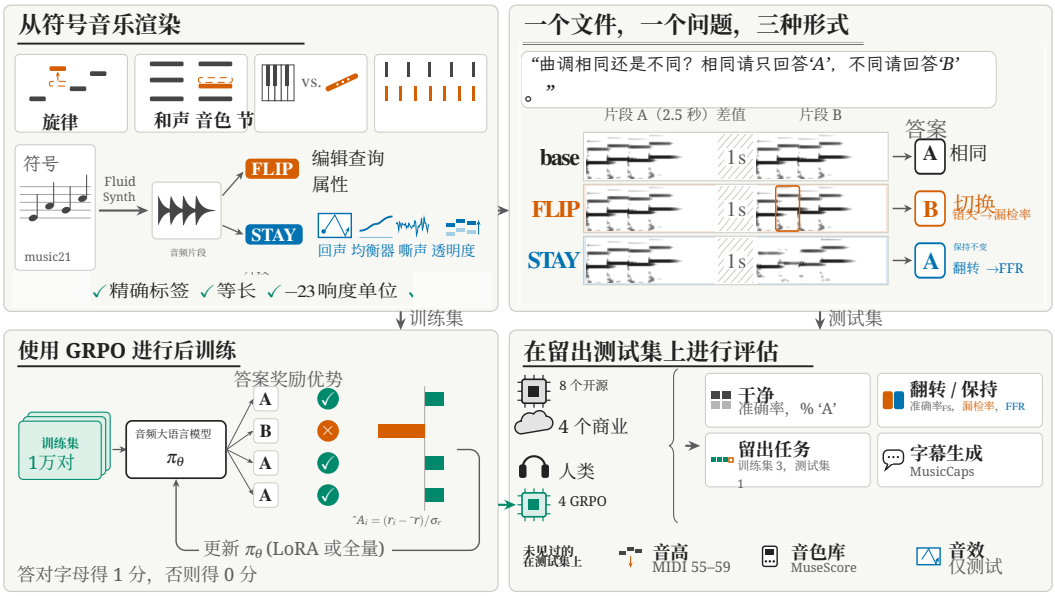

图 1：**MusicListenBench 概览**。题目由符号音乐渲染而成，因此每个答案都是确定的；另一个测试集则使用真实录音构建相同的题目（第 3.4 节）。翻转版本会改变查询属性和答案；保持版本只改变声音，答案保持不变。我们在留出条件下测试开放模型、商业音频大语言模型和人类听众，并使用干净配对上的 GRPO 对开放模型进行后训练。

表 1：四项评测任务。不同：当音频片段在查询属性上存在差异时，它们有何不同（对于节奏，即较快的音频片段有何不同）。音频片段：单个音频片段的时长。每项任务都有 2,500 个训练条目和 250 个干净测试项。

Task	问题	不同	级别	Clip
旋律	相同的旋律?	四音符动机中的一个音符发生移动	100、200、400、700 音分	2.5 秒
和声	相同的和弦性质?	和弦的三度音移动	100、200、300 音分	2.0 秒
音色	是相同的乐器吗?	乐器类别（钢琴、吉他、小提琴，长笛）	一个（分类）	2.5 秒
节奏	哪个音频片段更快?	踩镲音轨的敲击速率	速率比 1.10、1.20，1.35, 1.50	4.0 秒

完全相同，因此干净配对是测试集中最容易的版本。附录 A 给出了所有设置。对于节奏，我们省略起始音计数，因为数点击声本身就是任务。上述检查涵盖了文件内部的线索。两个音频片段之间关系中的线索，例如它们听起来有多相似，则是保持变体所测试的内容（第 3.3 节）。

3.3 翻转与保持：两种失败方式

对于 1,000 个干净测试项，我们通过替换片段 B，为每个测试项构建两个变体。

- 翻转：将片段 B 替换为编辑后的片段，使查询属性发生变化，例如移动一个音符、移位三度音或换成另一种乐器。对于节奏，则交换两个音频片段的位置。正确答案也随之切换。

- 保持：片段 B 替换为一段听起来不同、但保留查询属性的音频。我们采用四种此类变化：房间回声（混响时间为 0.2、0.5 或 1.0 秒，使用 pyroomacoustics 模拟；Scheibler 等，2018）、轻微的均衡器变化（在 200 Hz 和 4 kHz 处进行 ±3dB 架式均衡）、 −50 dBFS 的背景嘶声，以及仅针对旋律和和声、幅度最高为 200 音分的移调

。房间回声、均衡器变化和噪声也是音乐修复模型学习去除的退化类型（Melechovsky 等，2025）。正确答案保持不变。

移调测试相对音高。整段旋律整体升高后仍是同一旋律（Dowling，1978），但其中一个音符移动后就不再是了。四种保持变化均匀分布在各题目中。

我们报告两种错误率。漏检率（MR）是翻转项答错的比例：音乐发生变化后，模型仍保留原答案。误翻转率（FFR）是保持项答错的比例：只有声音发生变化时，模型却改变了答案。两者的平均值即为翻转/保持准确率，$\mathrm{Acc}_{\mathrm{FS}} = 1 - (\mathrm{MR} + \mathrm{FFR})/2$。这个分数有一个实用特性：同一题目的两个变体答案相反。在旋律、和声和音色题中，“A”始终表示相同，因此它们的正确字母也相反；所以，忽略音频的模型恰好能答对其中一个，得分正好是50%，无论它偏好哪个字母。节奏问题要求判断哪个音频片段更快或更慢，而且每个变体采用各自的措辞，因此250道节奏题中有131道的两个变体正确字母相同。在这些题目中，忽略音频的模型只能在平均意义上达到50%：始终回答“A”，节奏题得分为47.8%，总体得分为49.45%，而干净测试集上的得分为52.8%。

这两个比率能够捕捉不同的错误。在三个相同/不同任务中，对于一个只要两个音频片段听起来不同就回答“不同”的模型而言，它很少会漏判翻转变化，但对于答案为“相同”的每个保持版本音频对，它都会发生错误翻转。每个比率也会单独随字母偏差而变化（始终回答“A”会得到 53.0% 的漏检率和 48.1% 的错误翻转率），但两者的平均值仍接近 50%。错误翻转率会统计每个错误的保持版本答案，包括模型在干净音频对上已经犯下的错误。例如，对于一个在干净音频对存在差异时回答“相同”的模型，加入嘶声后，它很可能仍会回答“相同”，尽管嘶声并未改变任何内容。为了只统计新增错误，我们还报告 $\mathrm{FFR}_c = P$（保持版本错误 | 干净版本正确），即干净版本回答正确的题目上的错误翻转率。附录 M 给出了漏检的相同指标，以及一项仅当某个题目的干净、翻转和保持版本全部回答正确时才计入该题的得分。

3.4 数据集划分与留出条件

训练集划分包含 10,000 个题目（每项任务 2,500 个），每个题目都有仅声音版本（附录 A）。我们仅将其用于后训练。渲染测试集包含 1,000 个干净题目和 2,000 个翻转/保持项目。由同一生成项目生成的文件绝不会同时出现在两个数据集划分中。测试集与训练集还有三处不同。

- **音高范围。** 训练集中的旋律和音色题目以 MIDI 音符 60 至 72 为起点，而测试集中的则以 55 至 59 为起点。训练集中的和声根音取自 MIDI 48 至 60，测试集中的则取自 62 至 67。
- **音色库。** 训练音频使用 FluidR3 GM（Wen，2000年）或 GeneralUser GS（Collins，2024年）渲染，所有测试音频则使用 MuseScore General（Collins，2020年）渲染。
- **音效。** 房间回声、均衡器和嘶声仅出现在测试集中。我们的主要实验仅使用干净配对进行训练；一项对照实验加入了仅声音版本，但不加入房间回声、均衡器或嘶声（第5.3节）。

在测试项目中，字母的含义是固定的（A 表示相同），1,000 个答案中有 528 个是“A”。在训练项目中，我们对大约一半的项目交换字母的含义，因此字母永远无法预测答案。这四项任务还支持留一任务测试：在三项任务上进行训练，并在第四项任务上进行测试。

真实测试集。另一个测试集以相同方式从真实录音中构建，且从未用于训练（600 个干净项目和 1,200 个翻转/保持项目）。旋律和和声使用来自 GuitarSet（Xi 等，2018年）的吉他录音：我们对一个音符进行移调，或仅对演奏和弦三音的琴弦进行移调。音色使用来自 Medley-solos-DB（Lostanlen 等，2018年）的七种乐器的独奏录音，每对中的两个音频片段均来自不同的录音。节奏使用来自 MUSDB18-HQ（Rafii 等，2019年）的时间拉伸鼓声分轨。在旋律和和声的“相同”配对中，音频会以零音分通过移调器，因此两种配对都经过相同处理。（附录 Q）

3.5 协议与公开排行榜

MusicListenBench 配有公开排行榜[4]，，任何人都可以提交模型。网站会报告两个测试集各任务的干净准确率、翻转/保持准确率、漏检率、FFR，以及回答“A”的比例，因此字母偏差始终一目了然。零样本模型和使用我们训练集划分的模型会分别列出（附录 E）。由于渲染测试项目由生成器生成，如果公开测试集不再适用，我们可以使用新的随机种子生成一套全新的测试集。

4 实验设置

模型。 我们测试了八个开源音频大语言模型：Qwen2-Audio（Chu 等，2024年）、Qwen2.5-Omni（Xu 等，2025年）、Audio Flamingo 3（Ghosh 等，2026年）、Phi-4-multimodal（Abouelenin 等，2025年）、Kimi-Audio（Ding 等，2025年）、MiMo-Audio（小米，2025年）、Fun-Audio-Chat（Team 等，2025年）和 Step-Audio 2 mini（Wu 等，2025年）。我们还通过其 API 测试了四个商业模型：GPT-Audio-1.5（OpenAI，2026年）、GPT-Audio-mini（OpenAI，2026年）、Gemini 2.5 Pro（Comanici 等，2025年）和 Gemini 3.8 Flash（Google DeepMind，2026年）。对于开放模型，我们通过比较词元“A”和“B”的概率来读取答案；改为读取生成的字母时，每个条目得到的答案都相同。对于 API 模型，我们使用生成的字母。没有给出字母的回复，在表2中计为“A”，在真实测试集上则计为错误。附录 M 分别采用这两种方式对渲染测试进行评分；结论没有变化。附录 C 列出了所有检查点和 API 设置。

人类参考。 测试向所有人开放，不论其音乐背景如何。12 位听众，从休闲听众到有多年练习经验的音乐家，每人回答 24 道题：总计 288 个答案，每项任务 72 个，干净、翻转和保持测试各 96 个。他们通过耳机聆听了与模型相同的音频并回答了相同的问题（附录 B）。

GRPO。 我们使用 GRPO 对四种开放模型进行后训练：Qwen2-Audio、Qwen2.5-Omni、Audio Flamingo 3 和 Phi-4-multimodal。我们采用 LoRA（Hu 等，2021）或全量微调。奖励为答对时 1 分，否则为 0 分；此外，模型给出可解析选项时会获得少量奖励（0.2），另有一项更小的奖励（最多 0.1），鼓励模型选择其当前较少给出的字母，以防止字母塌缩（组大小为 8；LoRA 的学习率为 10^{-5}，全量微调的学习率为 10^{-6}；每步使用 8 个提示词，即 64 条采样轨迹；Qwen2.5-Omni 使用 LoRA 时则为 24 个提示词）。我们报告第 1,000 步的检查点；Audio Flamingo 3 全量微调的情况除外，其验证准确率在第 500 步达到峰值，之后模型对二元奖励过度优化（KL 值上升且出现熵塌缩），因此我们在该步停止训练。所有设置见附录 D。

留一任务法。 对于 Qwen2-Audio 和 Qwen2.5-Omni，我们在三个任务（7,500 个题目）上训练 LoRA GRPO 500 步，并在第四个任务的 250 个干净题目上进行测试。我们使用最终检查点，不查看测试集准确率来选择检查点。

字幕生成。 为检查是否存在副作用，我们让 Qwen2-Audio 和 Qwen2.5-Omni 在 GRPO 训练前后，对 MusicCaps 评估集划分中可用的 2,735 个音频片段生成字幕（Agostinelli 等，2023）；专用音乐字幕生成模型则针对这项任务进行训练（Chopra 等，2025）。对于每个模型，我们在相同的音频片段上使用 BLEU-4、METEOR、ROUGE-L 和 BERTScore-F1 将生成的字幕与参考字幕进行比较（Papineni 等，2002；Banerjee 与 Lavie，2005；Lin，2004；Zhang 等，2019）（附录 G）。

随机水平。 对于250个题目，得分在43.8%至56.2%之间时，与50%没有显著差异（双侧，$p > 0.05$，正态近似）；对于500、1,000和2,000个题目，这一范围分别为45.6%至54.4%、46.9%至53.1%和47.8%至52.2%。

表2：未经训练的结果（%）。干净：每项任务250个题目的准确率。翻转/保持：2,000个题目上的准确率（$\mathrm{Acc}_{\mathrm{FS}}$）、漏检率（MR）和误翻转率（FFR）；MR和FFR越低越好。粗体：每列中表现最佳的开放模型；带下划线：表现最佳的商业模型。最后两行是参考数据；人类行在每项任务中有24个干净答案，翻转和保持答案总计各96个。

	干净准确率					翻转/保持		
模型	Mel.	和声	Tim.	Rhy.	全部	$\mathrm{Acc}_{\mathrm{FS}}$	MR	FFR
开放模型								
Qwen2-Audio	46.0	50.4	43.2	49.2	47.2	50.7	47.7	50.9
Qwen2.5-Omni	**68.8**	**59.2**	**66.8**	50.0	**61.2**	**61.2**	**36.7**	**41.0**
Audio Flamingo 3	51.2	48.8	58.0	52.8	52.7	51.4	49.6	47.6
Phi-4-multimodal	54.4	47.2	56.8	50.8	52.3	49.1	53.5	48.3
Kimi-Audio	57.6	47.2	53.2	52.8	52.7	51.0	52.3	45.7
MiMo-Audio	54.0	54.4	58.0	**59.6**	56.5	52.2	49.4	46.1
Fun-Audio-Chat	50.0	51.6	60.0	49.6	52.8	53.6	46.6	46.1
Step-Audio 2	52.4	50.0	62.8	52.4	54.4	52.7	49.1	45.5
商业模型（*API*）								
GPT-Audio-mini	47.6	51.6	54.4	50.8	51.1	49.6	50.6	50.3
GPT-Audio-1.5	45.6	47.6	52.0	44.0	47.3	52.1	45.8	50.1
Gemini 2.5 Pro	80.0	68.8	88.4	79.6	79.2	74.2	24.7	26.9
Gemini 3.8 Flash	53.6	54.4	85.6	72.4	66.5	64.0	33.0	39.1
人类（12）	79.2	87.5	95.8	87.5	87.5	87.0	5.2	20.8
始终选择“A”					52.8	49.5	53.0	48.1

5 结果

5.1 音频大语言模型能听出不同了吗?

表2回答了标题中的问题：还没有。在干净配对这一最简单的测试集版本中，八个开放模型中有七个的得分介于47.2%和56.5%之间，接近始终回答“A”的52.8%。翻转/保持消除了偏好某个字母所带来的优势。五个开放模型的得分位于随机水平范围内（49.1%至52.2%），Fun-Audio-Chat和Step-Audio 2略高于该范围（分别为53.6%和52.7%），只有Qwen2.5-Omni明显高于该范围（61.2%）。两个GPT-Audio 模型的得分都位于随机水平范围内（49.6%和52.1%）。Gemini 3.8 Flash达到64.0%，Gemini 2.5 Pro达到74.2%，这是所有未经训练的模型中最高的得分，但仍比人类听众（87.0%；$p < 0.001$，Fisher精确检验）低12.8个百分点。

几种开放模型大多只给出同一个字母。Qwen2-Audio 对 98.6% 的干净题目回答“B”，而 Phi-4-multimodal 和 Audio Flamingo 3 分别对 95.9% 和 87.7% 的题目回答“A”（见图5及附录F中的表14）。如果对两个答案类别的准确率取平均，三者都处于随机水平（49.9%、49.7%和50.6%）。

按任务统计的干净得分同样带有这种偏差，因此我们根据翻转/保持进行任务比较（见附录F中的表15）。音色是大多数模型能够超过随机水平的领域：三个开放模型和三个商业模型在音色任务上的准确率都明显高于50%，但只有 Qwen2.5-Omni 和两个 Gemini模型在旋律任务上超过了这一水平。Gemini 3.8 Flash 在音色任务上达到84.2%，但在旋律和和声任务上分别只有57.2%和51.2%；我们的听众在这两项上的得分分别为85.4%和66.7%。Gemini 2.5 Pro 的表现更加均衡（旋律73.2%，和声63.8%），但在听众认为最容易的两个任务——音色（82.4%）和节奏（77.4%，而听众在两项上均为97.9%）——上的表现与我们的听众差距最大。它的大多数错误都是将存在不同的音频片段回答为“相同”：无论音符移动100还是700音分，它都只能检测出大约一半的移动音符；在36对音频片段中，有34对它将钢琴和吉他听成相同的乐器（见附录I）。

5.2 后训练能缩小差值吗?

GRPO 在八次运行中的七次提高了干净准确率，提升幅度为 10.0 至 37.8 个百分点（表3）。Qwen2.5-Omni 使用 LoRA 时达到 99.0%，使用全量微调时达到 96.2%。Audio Flamingo 3 达到 79.4%，Qwen2-Audio 达到 68.9%。测试音频使用的音高范围和音色库都与训练音频不同，因此这些提升并非来自对训练音频片段的记忆；对于 Qwen2.5-Omni，它们

[4]https://amaai-lab.github.io/MusicListenBench/

表3：GRPO 在留出测试集上的结果（%）。无：未进行训练。 †第500步；其他所有运行均为第1,000步。相同运行的字母偏差见图5和附录F。

		干净准确率					翻转/保持		
模型	GRPO	Mel.	Har.	Tim.	Rhy.	全部	Acc_{FS}	MR	FFR
Qwen2-Audio	none	46.0	50.4	43.2	49.2	47.2	50.7	47.7	50.9
	LoRA	70.8	60.8	92.4	51.6	68.9	63.8	29.7	42.7
	full	70.4	61.6	80.8	52.0	66.2	61.5	31.7	45.4
Qwen2.5-Omni	none	68.8	59.2	66.8	50.0	61.2	61.2	36.7	41.0
	LoRA	99.6	96.8	99.6	100.0	99.0	86.0	1.9	26.1
	full	100.0	85.2	100.0	99.6	96.2	83.5	4.2	28.9
Audio Flamingo 3	none	51.2	48.8	58.0	52.8	52.7	51.4	49.6	47.6
	LoRA	57.2	74.8	97.6	50.8	70.1	64.4	29.7	41.4
	全量†	68.0	80.8	99.6	69.2	79.4	70.4	22.5	36.7
Phi-4-multimodal	none	54.4	47.2	56.8	50.8	52.3	49.1	53.5	48.3
	LoRA	52.0	49.6	56.8	50.8	52.3	49.1	53.4	48.4
	full	55.6	54.0	88.8	50.8	62.3	58.9	42.9	39.3

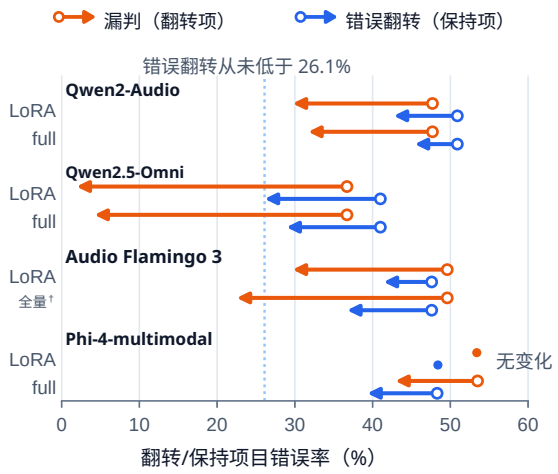

图2：GRPO 训练前后 flip/stay 任务中的漏判和错误翻转，每次运行占一行。每支箭头从未训练时的错误率（空心圆）出发，指向 GRPO 之后的错误率。橙色：漏检率（翻转项）；蓝色：错误翻转率（保持项）。在学会该任务的七次运行中，有六次的橙色箭头长度是蓝色箭头的 2.2 至 3.2 倍，而且这些运行的错误翻转率都未低于 26.1%（点线）。 †第500步。数值见表3。

也能迁移到真实录音中（第5.5节）。训练还减少了字母偏差：对于答案为“A”的题目和答案为“B”的题目，Audio Flamingo 3（全量）的准确率差值从75.3个百分点降至17.0个百分点，Qwen2-Audio（LoRA）的差值则从97.2个百分点降至39.6个百分点（附录F中的图5）。只有Qwen2.5-Omni（100.0%和99.6%）以及采用全量微调的Audio Flamingo 3（69.2%）学会了节奏任务；其他所有训练结果均介于50.8%和52.0%之间。

5.3 模型学会更多地注意变化，而不是忽略变化

干净准确率掩盖了 GRPO 教会模型的东西。图2展示了各模型的漏检率和错误翻转率如何随训练而变化。训练降低漏检率的幅度远大于降低错误翻转率的幅度。在七次取得学习效果的训练中，有六次漏检率的降幅是错误翻转率的 2.2 至 3.2 倍（$p < 0.01$ 每次训练；表16）。例外是采用全量微调的 Phi-4-multimodal，两种错误率的降幅相近（分别为 10.6 和 9.0 个百分点）。对于采用 LoRA 的 Qwen2.5-Omni，漏检率从 36.7% 降至 1.9%，但错误翻转率仅从 41.0% 降至 26.1%。训练后的模型如今错失的变化比 Gemini 2.5 Pro 少 13 倍（1.9%，而非 24.7%），但其错误翻转率并没有更低（26.1%，而非 26.9%）。不过，Gemini 2.5 Pro 的错误翻转中，大多数（旋律、和声和音色上的 209 次中有 155 次）其实是错失了音乐变化，而不是对声音作出反应。若只统计干净版本答对的题目（FFR_c），其错误翻转率会降至 20.8%，而训练后的模型仍为 25.7%（附录 M）。附录F中的图4将所有模型和我们的听众置于这两个坐标轴上。

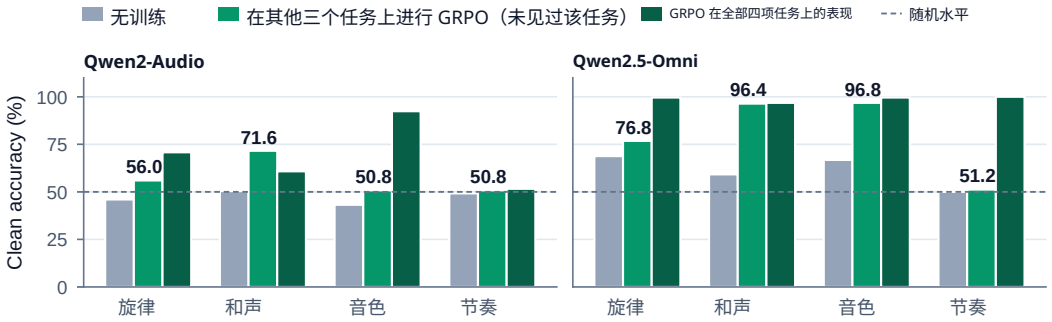

图 3：迁移到训练中未包含的任务（干净准确率，每个柱形代表 250 个条目）。灰色：未训练。绿色，带数值：在其他三项任务上使用 LoRA GRPO 训练 500 步，因此模型从未见过该任务本身。深绿色：在全部四项任务上使用 LoRA GRPO 训练 1,000 步。

人类听众的错误翻转也多于漏判（20.8%，而非 5.2%），但他们 20 次错误翻转中至少有 11 次发生在和声题目上（表9）；而在音色和节奏方面，他们对 48 个翻转/保持答案最多各错一次。训练后的 Qwen2.5-Omni 错失的变化少于我们的听众（1.9%，而非 5.2%），但在 500 个音色和节奏的保持不变配对中，有 94 个发生了错误翻转。

按声音变化统计的错误翻转。 每个保持不变的配对都应用四种声音变化之一，而我们的主要 GRPO 训练运行并未使用其中任何一种（附录 H 中的表 19）。使用 LoRA 的 Qwen2.5-Omni 学会了忽略均衡器，其错误翻转率从 33.7% 降至 1.4%，并且部分学会了忽略房间回声（从 37.2% 降至 16.6%）。但它没有学会忽略嘶声（训练前为 49.0%，训练后为 52.4%）或移调（48.4% 和 45.2%）。在音色和节奏任务上，与我们的听众之间的差值完全来自嘶声：嘶声造成了 94 次错误翻转中的 85 次；而在这些任务的其他 334 个保持不变的配对中（均衡器或房间回声），模型错了 9 次（2.7%）。模型还能够在 99.6% 的旋律变化配对中检测出单个移动音符，但在仅发生移调的旋律配对中仍有 40.3% 判断错误（训练前为 46.8%）。它能听出音高发生了变化，却还没有学会相对音高。全量微调、Qwen2-Audio 和 Audio Flamingo 3 呈现出相同的模式（附录 H 中的图 6）。

即使在训练之前，这种排名也成立。未经 GRPO 训练时，Qwen2.5-Omni 和 Gemini 3.8 Flash 在均衡器和嘶声上的错误最少，在嘶声和移调上的错误最多；而我们未训练的 Gemini 2.5 Pro 在移调和嘶声上的错误最多（分别为 46.8% 和 32.5%，相比之下，均衡器为 22.4%，房间回声为 17.2%）。因此，对嘶声的弱点并非由我们的训练造成：GRPO 会强化这一排名，但不会改变它。奖励只检查干净配对上的字母，因此模型获得的任何不变性都是副作用。

错误翻转从何而来。每次运行中，音频编码器都处于冻结状态，因此训练只会改变语言模型读取其输出的方式。在我们渲染的音频片段中，嘶声导致的输出变化幅度是真实音乐变化的 2.75 倍，而均衡器导致的变化幅度仅为 0.04 倍（附录L）。这一距离可以预测训练后的模型会将哪些“相同”保持不变的配对判为不同（曲线下面积为 0.92，训练前为 0.74）。训练教会模型的还不只是针对这一距离设置阈值，否则均衡器和房间回声造成的错误翻转会多得多。编码器输出还保留了忽略嘶声所需的大部分信息：在线性探针上用房间回声、均衡器和嘶声示例进行训练后，嘶声造成的错误翻转率从 55.0% 降至 32.5%。

训练算法与数据。 对相同的干净配对进行监督微调，结果也呈现相同趋势：漏判减少了 33.7 个百分点，错误翻转减少了 13.2 个百分点（附录 K）。在 GRPO 中加入训练集划分的仅声音版本（编解码器、音色库替换和移调）后，渲染片段在房间回声下的错误翻转率从 16.6% 降至 4.5%，在嘶声下则从 52.4% 降至 37.7%，尽管这两种变化都没有出现在训练中。翻转/保持准确率从 86.0% 上升至 88.2%。在移调条件下，错误翻转率仍为 45.2%，尽管移调音对现已纳入训练集。

5.4 迁移到未见任务及其副作用

迁移到未见任务。在三个任务上训练可以提升第四个任务，呈现出清晰的规律（图 3）。两个模型的和声表现都有提升。Qwen2-Audio 从 50.4% 提升至 71.6%，高于

和声包含在训练集中时达到的 60.8%。Qwen2.5-Omni 从 59.2% 提升至 96.4%，接近和声包含在训练集中时的 96.8%。Qwen2.5-Omni 的音色表现有所提升（从 66.8% 提升至 96.8%），旋律表现则提升了 8 到 10 个百分点。对 Qwen2-Audio 而言，音色表现仅从 43.2% 上升至 50.8%，仍处于随机水平。节奏表现没有提升（仅提高 1.2 和 1.6 个百分点），尽管节奏包含在训练集中时，Qwen2.5-Omni 在节奏任务上达到了 100.0%。节奏是唯一涉及时间的任务，也是唯一一个不是“相同/不同”问题的任务。因此，模型似乎学会了一种通用的“相同/不同”技能，能够帮助完成任何此类任务，但无法判断哪个音频片段更快。

对字幕生成的副作用。 在合成配对上进行 GRPO，能使字幕生成表现接近基底模型（附录 G 中的表17）。在 2,735 个 MusicCaps 音频片段上，BERTScore-F1 变化了 -1.28 至 $+1.95$ 个百分点，BLEU-4 变化了 -0.22 至 $+1.73$。平均字幕长度变化不到一个词，且没有一次运行生成退化的字幕（配对测试见附录 G）。

5.5 真实录音

开放模型的表现接近随机水平。 在真实测试集上，八个开放模型在翻转/保持任务中的得分为 48.3% 至 55.5%，只有 Audio Flamingo 3 和 Fun-Audio-Chat 的表现显著高于随机水平（表39）。Gemini 2.5 Pro 再次成为表现最好的未训练模型（73.8%；在相同评分标准下，其在渲染测试中的得分为 73.7%）。在全部 12 个未训练模型中，排名与渲染测试一致（斯皮尔曼 $\rho = 0.80, p = 0.002$）。

训练效果得以迁移。 在我们合成的配对上训练后，Qwen2.5-Omni 的准确率从 52.1% 提升至 62.8%（精确麦克尼马尔检验，$p < 0.001$）。监督微调后的准确率为 62.9%，加入仅声音配对后则达到 70.1%。其他训练后的模型与随机水平相差不超过 6 个百分点。在 Slakh2100 配对上训练（附录 J）后，Qwen2.5-Omni 的准确率达到 75.0%，与 Gemini 2.5 Pro 持平（$p = 0.50$）。

同一和弦的两次扫弦。 同一和弦的两次扫弦总会略有不同，但听者应该将它们判为相同（表41）。训练后的 Qwen2.5-Omni 能识别出 41 组移调三度音程中的 38 组，但也会将 42 组此类同和弦配对中的 27 组判为不同。正如第5.3节所述，它往往会把两段录音之间的任何差异都视为变化。训练中加入仅声音配对后，它只将 42 组中的 7 组判为不同（$p < 0.001$），并识别出 41 组移调三度音程中的 26 组。Gemini 2.5 Pro 则容易犯相反的错误：它会将 42 组同和弦配对中的 37 组判为相同，但只能识别出 41 组移调三度音程中的 12 组。

真实音频中的错误翻转。 训练后，在全部三个测试集上，Qwen2.5-Omni 的每次运行中，错误翻转都保持在与漏判相当或更高的水平。在真实录音中，模型也会漏判更多变化，因此两项比率最终非常接近（37.0% 和 37.3%）；对于所有开放模型而言，真实旋律都很难（干净准确率最高仅为 55.3%，而 Gemini 2.5 Pro 为 72.7%）。在三种音效中，嘶声在渲染片段中造成的错误翻转最多（52.4%），但在真实录音中最少（20.9%，而均衡器为 46.6%），可能是因为真实录音中本身已经含有一些噪声。在全部三个测试集上，每次训练后的 Qwen2.5-Omni 运行仍会将 38% 至 49% 的移调音对判断错误。

进一步检查。 附录 M 至 P 增加了配对指标和逐任务测试、将每个编辑后的片段播放两次的对照实验、12 音符和 24 音符的旋律，以及更细致的音色测试。对于更长的旋律，除 Qwen2.5-Omni 外，未经训练的开放模型仍接近随机水平（48.5% 至 56.8%），而训练后模型的错误翻转仍然很高。对于音色接近的乐器对，训练后的模型只能检测出 10% 的小提琴/中提琴变化。

6 讨论

干净准确率还不够。 在我们的干净测试集上，始终回答“A”可得 52.8%；在翻转/保持测试中则为 49.45%。GRPO 将 Qwen2.5-Omni 在干净配对上的准确率提升至 99.0%，但在翻转/保持测试中仅为 86.0%；监督微调则分别提升至 97.4% 和 84.6%。干净准确率无法区分真正比较音乐的模型与对任何声音差异都有反应的模型，而配对变体可以，因此基准测试应同时报告这两项指标。这不仅关系到基准测试：音频大语言模型已被用于为音乐添加标签、筛选和排序；如果评判器在加入轻微嘶声后就改变判定，它评估的便是录音，而非音乐。我们测试中的每段音频都只包含一种乐器或一条音轨，因此尚未测试完整混音。

7 结论

通过精确标签和成对的翻转/保持变体，MusicListenBench 表明，大多数开源音频大语言模型无法听出两段音乐片段之间的差异，无论音频是渲染的还是真实的；而表现最佳的商业模型 Gemini 2.5 Pro 仍比人类听众低 12.8 个百分点。经过少量 GRPO 训练后，最佳模型在翻转/保持测试中的表现至多与我们的听众持平：它学会了注意变化，并且在渲染片段中忽略轻微的均衡器变化，但无法忽略轻微嘶声或移调。仅用干净配对进行训练，从未要求模型忽略声音变化，因此必须明确要求模型具备不变性，而这项要求的代价很低：奖励只需规定，对于某个条目的干净版本和保持版本（两者来自同一条目），模型必须给出相同答案。在训练中加入仅声音配对已有帮助，但仍无法教会模型相对音高。若模型从未接受过保持类变化的训练，却能在翻转/保持测试中达到我们听众的区间（82.3% 至 91.1 %；附录 B），并且像听众一样，几乎不会因音色和节奏而错误翻转，那么我们就认为标题中的问题已有了答案。训练后的 Qwen2.5-Omni 满足第一个条件（86.0%），但未满足第二个条件（错误翻转率分别为 20.8% 和 16.8%）。教会模型哪些差异应该忽略，仍是一个有待解决的问题。

大语言模型使用声明

我们使用了大语言模型助手（Claude Sonnet）改写句子并进行语法检查。所有研究构想、实验和结论均由作者提出，作者对内容承担全部责任。

伦理声明

人类听众。 共有十二名成年人参加了听力测试。他们事先获知任务内容、参与完全自愿且可以随时退出，并在开始前提供了知情同意。我们只记录了答案以及对音乐经验的粗略自我报告。我们没有收集姓名、联系方式、音频录音或其他身份信息，因此公开发布的答案无法追溯到具体参与者。听众在测试开始前自行设置了舒适的播放音量。

音频与第三方材料。 渲染集中的所有音频均由生成的符号音乐通过免费提供的音源库合成，因此不包含任何人的录音或受版权保护的音乐作品。我们遵循各音源库的许可条款（Wen，2000年；Collins，2024年；2020年），并在发布时注明这些条款。MusicCaps 音频片段仅用于衡量字幕生成的副作用。我们不会重新分发这些音频片段；我们发布字幕和分数，音频片段仍保留在其原始来源处。真实测试集（第 3.4 节）和 Slakh2100 研究（附录 J）使用遵循各自许可协议的公开研究数据集：GuitarSet（知识共享署名-非商业性使用 4.0）、Medley-solos-DB（知识共享署名 4.0）、MUSDB18-HQ（研究用途，申请后可访问）和 Slakh2100（知识共享署名 4.0）。与渲染集不同，这些数据集包含表演者的录音和现有歌曲。对于真实测试集，我们发布项目列表和构建脚本，但不发布任何 MUSDB18-HQ 音频。

可复现性声明

排行榜可访问 https://amaai-lab.github.io/MusicListenBench/。我们在 https://github.com/ AMAAI-Lab/MusicListenBench 发布了生成器、渲染集的全部音频和项目文件、真实测试集的项目列表和构建脚本、评分脚本以及评估代码；数据集托管于 https://huggingface.co/ datasets/amaai-lab/MusicListenBench。

附录 A 给出了全部生成参数，附录 C 给出了全部模型版本，附录 D 给出了全部训练设置。

参考文献

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen 等。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*，2025年。Andrea Agostinelli、Timo I. Denk、Zal´an Borsos、Jesse Engel、Mauro Verzetti、Antoine Caillon、Qingqing Huang、Aren Jansen、Adam Roberts、Marco Tagliasacchi、Matt Sharifi、Neil Zeghidour 和 Christian Frank。MusicLM：从文本生成音乐。*arXiv* 预印本 *arXiv:2301.11325*，2023年。Satanjeev Banerjee 和 Alon Lavie。METEOR：一种自动机器翻译评估指标，与人类判断的相关性更高。载于*ACL* 机器翻译和/或摘要生成内在与外在评估指标研讨会论文集，第 65–72 页，2005 年。Brandon James Carone 、 Iran R Roman 和 Pablo Ripoll´es。评估多模态大语言模型在核心音乐感知任务上的表现。*arXiv* 预印本 *arXiv:2510.22455*，2025年。Brandon James Carone、Iran R Roman 和 Pablo Ripoll´es。MUSE 基准：探究音频大语言模型的音乐感知与听觉关系推理。载于*ICASSP 2026-2026 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 16097–16101 页。IEEE，2026年。Anuradha Chopra、Abhinaba Roy 和 Dorien Herremans。Sonicverse：用于音乐特征引导的字幕生成的多任务学习。*arXiv* 预印本 *arXiv:2506.15154*，2025年。Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、Yichong Leng、Yuanjun Lv、Jinzheng He、Junyang Lin 等。Qwen2-Audio 技术报告。*arXiv* 预印本 *arXiv:2407.10759*，2024年。S. Christian Collins。MuseScore General：MuseScore 默认音色库。https:// musescore.org/en/node/317991，2020年。源自 FluidR3Mono，采用 MIT 许可证；访问日期：2026年9月20日。S. Christian Collins。GeneralUser GS：通用 MIDI 音色库。https://github. com/mrbumpy409/GeneralUser-GS，2024年。访问日期：2026年9月20日。Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen 等。Gemini 2.5：凭借高级推理、多模态能力、长上下文和新一代智能体能力拓展前沿。*arXiv* 预印本 *arXiv:2507.06261*，2025年。Michael Scott Cuthbert 和 Christopher Ariza。music21：用于计算机辅助音乐学和符号音乐数据的工具包。2010年。Ding Ding、Zeqian Ju、Yichong Leng、Songxiang Liu、Tong Liu、Zeyu Shang、Kai Shen、Wei Song、Xu Tan、Heyi Tang 等。Kimi-Audio 技术报告。*arXiv* 预印本 *arXiv:2504.18425*，2025年。W Jay Dowling。音阶与轮廓：旋律记忆理论的两个组成部分。心理学评论，85(4):341，1978年。Milan Liessens Dujardin、Song-Ze Yu、Craver Corbyn Thomas-Smith、David M Chan 和 Ka- rina Nguyen。PitchBench：测量音频语言模型的音高听辨能力。*arXiv* 预印本 *arXiv:2605.26176*，2026年。R EBU-Recommendation。音频信号的响度标准化与允许的最大电平。欧洲广播联盟，2011年。FluidSynth 开发者。FluidSynth：实时 SoundFont 合成器。https://www. fluidsynth.org，2000年。访问日期：2026年9月20日。Matt Gardner、Yoav Artzi、Victoria Basmov、Jonathan Berant、Ben Bogin、Sihao Chen、Pradeep Dasigi、Dheeru Dua、Yanai Elazar、Ananth Gottumukkala 等。通过对比集评估模型的局部决策边界。载于计算语言学协会研究成果：*EMNLP 2020*，第 1307–1323 页，2020年。

Sreyan Ghosh、Arushi Goel、Jaehyeon Kim、Sonal Kumar、Zhifeng Kong、Sang-gil Lee、Chao-Han Yang、Ramani Duraiswami、Dinesh Manocha、Rafael Valle 等。《Audio Flamingo 3：借助完全开源的大型音频语言模型推进音频智能》。神经信息处理系统进展，38:41819–41886，2026年。

Google DeepMind。Gemini 3.8 Flash 模型说明卡。https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-8-Flash-Model-Card.pdf，2026年。访问日期：2026年9月26日。

Michael J Hautus、Neil A Macmillan 和 C Douglas Creelman。检测理论：用户指南。Routledge，2021年。Dorien Herremans 和 Abhinaba Roy。使生成式音乐人工智能符合人类偏好：方法与挑战。载于*AAAI*人工智能大会论文集，第40卷，第39699–39706页，2026年。Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大型语言模型的低秩适配。*arXiv* 预印本 *arXiv:2106.09685*，2021年。Yonghyun Kim、Wayne Chi、Anastasios Angelopoulos、Wei-Lin Chiang、Koichi Saito、Shinji Watan- abe、Yuki Mitsufuji 和 Chris Donahue。Music Arena：文本生成音乐的实时评测。神经信息处理系统进展，第38卷，2026年。Lily NC Law 和 Marcel Zentner。客观评估音乐能力：音乐感知技能档案的构建与验证。*PLOS ONE*，7(12):e52508，2012年。Baiqi Li、Zhiqiu Lin、Wenxuan Peng、Jean De Dieu Nyandwi、Daniel Jiang、Zixian Ma、Simran Khanuja、Ranjay Krishna、Graham Neubig 和 Deva Ramanan。NaturalBench：在自然对抗样本上评估视觉语言模型。神经信息处理系统进展，第37卷，第17044–17068页，2024年。Gang Li、Jizhong Liu、Heinrich Dinkel、Yadong Niu、Junbo Zhang 和 Jian Luan。强化学习优于监督微调：音频问答案例研究。*arXiv* 预印本 *arXiv:2503.11197*，2025年。Chin-Yew Lin。ROUGE：用于自动评估摘要的工具包。载于文本摘要拓展，第74–81页，2004年。Fernando L´opez、Santosh Kesiraju 和 Jordi Luque。大型音频语言模型在选择题评估中的鲁棒性评估。*arXiv* 预印本 *arXiv:2510.04584*，2025年。Vincent Lostanlen、Carmine-Emanuele Cella、Rachel Bittner 和 Slim Essid。Medley-solos-DB：用于乐器识别的跨集合数据集。*Zenodo*：梅兰，瑞士，2018年。Tongyu Lu、Charlotta-Marlena Geist、Jan Melechovsky、Abhinaba Roy 和 Dorien Herremans。MelodySim：衡量旋律感知音乐相似度以进行抄袭检测。*arXiv* 预印本 *arXiv:2505.20979*，2025年。Yinghao Ma、Siyou Li、Juntao Yu、Emmanouil Benetos 和 Akira Maezawa。CMI-Bench：用于评估音乐指令遵循能力的综合基准测试。*arXiv* 预印本 *arXiv:2506.12285*，2025年。Ziyang Ma、Yinghao Ma、Yanqiao Zhu、Chen Yang、Yi-Wen Chao、Ruiyang Xu、Wenxi Chen、Yuanzhe Chen、Zhuo Chen、Jian Cong 等。MMAR：用于深度推理的挑战性基准测试，涵盖语音、音频、音乐及其混合内容。神经信息处理系统进展，第38卷，2026年。

Ethan Manilow、Gordon Wichern、Prem Seetharaman 和 Jonathan Le Roux。削减音乐源分离中的 Slakh：用于研究训练数据质量和数量影响的数据集。载于*2019*年*IEEE*音频与声学信号处理应用研讨会（*WASPAA*），第45–49页。IEEE，2019年。

Stephen McAdams、Suzanne Winsberg、Sophie Donnadieu、Geert De Soete 和 Jochen Krimphoff。合成音乐音色的知觉标度：共同维度、特异性与潜在受试者类别。心理学研究，58(3)：177–192，1995年。

Jan Melechovsky、Ambuj Mehrish、Abhinaba Roy 和 Dorien Herremans。Sonicmaster：迈向可控的一体化音乐修复与母带处理。*arXiv* 预印本 *arXiv:2508.03448*，2025年。

OpenAI。GPT-Audio-1.5 模型文档。https://developers.openai.com/api/docs/models/gpt-audio-1.5，2026年。访问日期：2026年9月20日。

OpenRouter 团队。Openrouter：面向人工智能模型的统一 API。https://openrouter.ai，2026年。访问日期：2026年9月26日。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第*40*届年会论文集，第311–318页，2002年。

ZafarRafii、AntoineLiutkus、Fabian-Robert St¨oter、Stylianos Ioannis Mimilakis 和 Rachel Bittner。MUSDB18-HQ——MUSDB18 的未压缩版本，2019年8月。网址：https://doi.org/ 10.5281/zenodo.3338373。

Marco Tulio Ribeiro、Tongshuang Wu、Carlos Guestrin 和 Sameer Singh。《超越准确率：使用 CheckList 对自然语言处理模型进行行为测试》。载于计算语言学协会第 *58* 届年会论文集，第 4902–4912 页，2020年。

Andrew Rouditchenko、Saurabhchand Bhati、Edson Araujo、Samuel Thomas、Hilde Kuehne、Rogerio Feris 和 James Glass。《Omni-R1：微调音频大语言模型真的需要音频吗？》载于*2025*年 *IEEE* 自动语音识别与理解研讨会（*ASRU*），第 1–7 页。IEEE，2025年。

Abhinaba Roy、Junyi Liang 和 Dorien Herremans。《MERIT：学习用于音频相似度的解耦音乐表征》。*arXiv* 预印本 *arXiv:2605.27346*，2026a。

Abhinaba Roy、Geeta Puri 和 Dorien Herremans。《Text2MIDI-InferAlign：通过推理时对齐改进符号音乐生成》。载于*ICASSP 2026-2026 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 14817–14821 页。IEEE，2026b。

Sakshi Sakshi、Utkarsh Tyagi、Sonal Kumar、Ashish Seth、Ramaneswaran Selvakumar、Oriol Nieto、Ramani Duraiswami、Sreyan Ghosh 和 Dinesh Manocha。MMAU：大规模多任务音频理解与推理基准测试。载于国际学习表征会议，第 2025 卷，第 84929–84964 页，2025年。

Robin Scheibler、Eric Bezzam 和 Ivan Dokmani´c。Pyroomacoustics：用于音频房间模拟和阵列处理算法的 Python 软件包。载于*2018*年 *IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 351–355 页。IEEE，2018年。

Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、YK Li、Yang Wu 等。DeepSeekMath：突破开放语言模型数学推理的极限。*arXiv* 预印本 *arXiv:2402.03300*，2024年。

Yirong Sun、Yanjun Chen、Xin Qiu、Gang Zhang、Hongyu Chen、Daokuan Wu、Chengming Li、Min Yang、Dawei Zhu、Wei Zhang 等。SonicBench：剖析大型音频语言模型中的物理感知瓶颈。*arXiv* 预印本 *arXiv:2601.11039*，2026年。

通义Fun团队、Qian Chen、Luyao Cheng、Chong Deng、Xiangang Li、Jiaqing Liu、Chao-Hong Tan、Wen Wang、Junhao Xu、Jieping Ye 等。Fun-Audio-Chat 技术报告。*arXiv* 预印本 *arXiv:2512.20156*，2025年。

Tristan Thrush、Ryan Jiang、Max Bartolo、Amanpreet Singh、Adina Williams、Douwe Kiela 和 Candace Ross。Winoground：探究视觉语言模型的视觉语言组合性。载于*2022*年 *IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，第 5228–5238 页。IEEE，2022年。

Shengbang Tong、Zhuang Liu、Yuexiang Zhai、Yi Ma、Yann LeCun 和 Saining Xie。眼睛大睁却视而不见？探索多模态大语言模型的视觉缺陷。载于 *2024*年 *IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，第 9568–9578 页。IEEE，2024年。Jean-Marc Valin、Koen Vos 和 Timothy Terriberry。Opus音频编解码器的定义。技术报告，2012年。Mikkel Wallentin、Andreas Højlund Nielsen、Morten Friis-Olivarius、Christian Vuust 和 Peter Vuust。音乐耳测试：一种用于测量音乐能力的新型可靠测试。《学习与个体差异》，20(3):188–196，2010年。Alex Warstadt、Alicia Parrish、Haokun Liu、Anhad Mohananey、Wei Peng、Sheng-Fu Wang 和 Samuel R Bowman。BLiMP：英语语言最小对基准。《计算语言学协会会刊》，8:377–392，2020年。Benno Weck、Ilaria Manco、Emmanouil Benetos、Elio Quinton、George Fazekas 和 Dmitry Bogdanov。MuChoMusic：评估多模态音频语言模型的音乐理解能力。*arXiv* 预印本 *arXiv:2408.01337*，2024年。Frank Wen。FluidR3 GM：一种通用 MIDI 音色库。https://musical-artifacts. com/artifacts/738，2000年。MIT 许可证；访问日期：2026年9月20日。Boyong Wu、Chao Yan、Chen Hu、Cheng Yi、Chengli Feng、Fei Tian、Feiyu Shen、Gang Yu、Haoyang Zhang、Jingbei Li 等。Step-Audio 2 技术报告。*arXiv* 预印本 *arXiv:2507.16632*，2025年。Qingyang Xi、Rachel M Bittner、Johan Pauwels、Xuzhou Ye 和 Juan Pablo Bello。GuitarSet：一个吉他转录数据集。第 19 届国际音乐信息检索学会会议（ISMIR）论文集，2018年。LCT Xiaomi。MiMo-Audio：音频语言模型是少样本学习器。*CoRR，abs/2512.23808*，2025年。Jin Xu、Zhifang Guo、Jinzheng He、Hangrui Hu、Ting He、Shuai Bai、Keqin Chen、Jialin Wang、Yang Fan、Kai Dang、Bin Zhang、Xiong Wang、Yunfei Chu 和 Junyang Lin。Qwen2.5-Omni 技术报告。*arXiv* 预印本 *arXiv:2503.20215*，2025年。Qian Yang、Jin Xu、Wenrui Liu、Yunfei Chu、Ziyue Jiang、Xiaohuan Zhou、Yichong Leng、Yuanjun Lv、Zhou Zhao、Chang Zhou 等。AIR-Bench：通过生成式理解对大型音频语言模型进行基准测试。载于 第 *62* 届计算语言学协会年会论文集（第 *1* 卷：长篇论文），第 1979–1998 页，2024年。Yongyi Zang、Sean O'Brien、Taylor Berg-Kirkpatrick、Julian McAuley 和 Zachary Novack。你真的在听吗？提升音乐问答基准中的感知意识。*arXiv* 预印本 *arXiv:2504.00369*，2025年。Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q Weinberger 和 Yoav Artzi。BERTScore：使用 BERT 评估文本生成。*arXiv* 预印本 *arXiv:1904.09675*，2019年。Chujie Zheng、Hao Zhou、Fandong Meng、Jie Zhou 和 Minlie Huang。大型语言模型并非稳健的多项选择器。载于 国际学习表征会议，第 2024 卷，第 19426–19454 页，2024年。

基准构建

生成。 表 4 给出了生成器设置。所有设置（包括划分规则）都已在训练任何模型之前固定在配置文件中。

Tabl图 4：生成器设置。MIDI 音符 60 为中央 C。训练集和测试集的音高范围不重叠 .

Task	设置
旋律	四音符动机，每个音符持续 0.4 秒，置于 2.5 秒的音频片段中。旋律不同时，其中一个音符会移动 100、200、400 或 700 音分。起始音高在 MIDI 60 至 72 之间（训练集），或在 55 至 59 之间（测试集）。
和声	2.0 秒音频片段中的三和弦。和弦性质不同时，三度音会移动 100、200 或 300 音分。大三度音向下移动，小三度音向上移动，因此移动 100 音分会将大调变为小调，将小调变为大调。移位三度音与根音及五度音之间至少相差 50 音分。根音从 MIDI 48 到 60（训练集），或从 62 到 67（测试集）。
音色	在 2.5 秒音频片段中，由钢琴、吉他、小提琴或长笛演奏的短旋律。两种不同的乐器总是来自不同的乐器家族；原声钢琴和电钢琴等相近配对不在此处纳入，而是在附录 P 中进行测试。起始音高与旋律相同。
节奏	4.0 秒音频片段中的踩镲节拍轨，每次点击持续 50 毫秒。两个音频片段的节拍速率比率差异为 1.10、1.20、1.35 或 1.50。第一次点击的时间独立于答案随机确定。

每个条目的音频版本。我们从每个生成项目中渲染出一个基底配对（音频片段 A 和 B）、一个会改变答案的编辑后的片段 B，以及两个或三个仅改变声音的片段 B。表 5 列出了仅声音变化。这些变化均保持查询属性不变，因此正确答案保持不变。与标准的相同/不同设计一样，旋律、和声和音色的干净“相同”配对中的音频片段 A 和 B 完全相同；保持版本仅在声音上为这些相同配对引入差异（第 3.3 节）。

表 5：按数据集划分和任务列出的仅声音变化。

变化	设置	数据集划分	任务
房间回声	pyroomacoustics 模拟、混响时间 0.2, 0.5 或 1.0 秒	test	all
EQ	搁架滤波器，在 200 Hz 和 4 kHz 处为 ± 3 dB	test	all
Hiss	添加了 -50dBFS 的噪声	test	all
编解码器	Opus（48 kbps，Valin等，2012年）或 MP3（128 kbps）	训练集	all
音色库替换	相同乐器，其他音色库	训练集	旋律、和声、节奏
新旋律、调	重新绘制的旋律和调，相同乐器	训练集	音色
移调	0 至 200 音分	训练集、测试集	旋律、和声

时长、响度和格式。每次更改后，我们都会进行解码（针对编解码器）、裁剪，并检查采样数量是否等于任务的片段时长。因此，对于旋律和音色，测试文件（片段 A、1.0 秒静音、片段 B）的时长为 6.0 秒；对于和声，时长为 5.0 秒；对于节奏，时长为 9.0 秒，并且任务中的每个文件都具有相同的采样数量。每个片段都归一化为 -23 LUFS（欧洲广播联盟建议，2011 年），并以 44.1 千赫、单声道的 16 位 PCM WAV 格式存储。

翻转和保持题目。这些题目没有使用新的音频，而是重新组合测试渲染。每个干净测试条目对应一个翻转项和一个保持项。每个保持项都保留原始片段 A，并从某个仅声音变化中选取片段 B。对于旋律、和声和音色，编辑只会改变片段 B，编辑后渲染结果中的片段 A 与原始片段 A 字节完全相同。节奏编辑会交换两个片段，因此节奏翻转项使用交换后的片段对。每个翻转项的正确答案都不同于其基础项（1,000 项中有 1,000 项）。对于旋律、和声和音色，正确字母也会改变。对于节奏，较快的片段会改变，但每个版本都会各自选用一种表述（较快或较慢），因此 250 个节奏翻转项中有 130 个的正确字母发生改变（第 3.3 节）。仅声音变化会在各题目间轮换（表 6）。

问题。 每项任务包含 8 到 10 种手写题目表述。节奏任务包含 10 种题目表述，其中一些要求选择更快的音频片段，另一些要求选择更慢的音频片段，正确字母取决于题目表述。每个

表 6：翻转/保持测试的组成（2,000 个题目）。

		保持变化				
Task	FLIP	EQ	Hiss	Room	移调	总计
旋律	250	63	63	62	62	500
和声	250	63	63	62	62	500
音色	250	84	83	83	n/a	500
节奏	250	84	83	83	n/a	500
总计	1,000	294	292	290	124	2,000

问题都以一条固定指令结尾，说明每个字母的含义，例如“仅在相同的情况下回答‘A’，不同的情况下回答‘B’.....”每个测试文件使用固定种子抽取一种题目表述，因此某个条目的基底、翻转和保持版本通常具有不同的题目表述。测试项目保持固定的字母含义。训练项目会对大约一半的题目交换字母含义，并进行平衡，使字母不携带任何关于答案的信息。表7和表8列出了两种字母含义下的每一种题目表述及其精确指令。

表 7：全部 34 条问题措辞，逐字呈现。每个提示均由题目表述 +固定指令（表8）以及音频占位符<音频>组成，条目的拼接音频片段附加在该占位符处。标准答案字母经过平衡（训练集：49.6% 为 A；测试集：52.8% 为 A）。

Task	问题措辞	正确字母是 the
旋律	两段旋律片段之间有一段短暂的静音。它们演奏的是相同的旋律，还是不同的旋律?	–
	你将听到两段旋律，中间有一段短暂停顿。两段旋律相同还是不同?	–
	聆听两个旋律短句，它们之间有一个短暂的差值。相同曲调，还是不同?	–
	一对旋律前后紧接着呈现，中间有一段短暂的静音。它们是相同的旋律吗?	–
	两个音乐乐句在短暂的停顿后依次出现。是否第二个乐句的旋律与第一个相同，还是不同?	–
	给定两段旋律素材音频片段，中间有一小段静音间隔。它们的旋律相同吗?	–
	仔细聆听由短暂的静音分隔的两段旋律音频片段。旋律相同，还是不同?	–
	依次播放两段短曲调，中间有一段停顿它们。听到的是同一段旋律重复两次，还是两段不同的旋律?	–
和声	依次弹奏两个和弦，中间有一段短暂的静音它们。两个和弦相同吗，还是不同?	–
	你将听到两个和弦，中间隔着一段短暂停顿。相同和弦，还是不同?	–
	听一听由一小段静音隔开的两个和弦。它们听起来像是同一种和弦，还是不同?	–
	一对和弦接连播放，中间有一小段静音在两者之间。和声相同还是不同?	–
	短暂停顿后，两个短和弦依次出现。相同和弦类型，还是不同?	–

下页继续

表 7（续）

Task	问题措辞	正确字母是 the
	给你两个中间隔着一小段静音间隔的和弦片段。它们是否共享相同的和弦类型?	–
	听听这两个和弦，中间有一小段静音。相同和声性质，还是不同?	–
	依次演奏两个和弦，中间有一段停顿它们。是同一个和弦演奏了两次，还是两个不同的和声性质?	–
音色	两个音频片段之间插入了一小段静音。它们是由相同的乐器 , 还是不同的乐器?	–
	你将听到两段由短暂停顿隔开的短音频片段。相同两者中的乐器 , 还是不同?	–
	听两段中间有短暂间隔的音频片段。它们是相同的乐器 , 还是不同的乐器?	–
	连续播放两段音频片段，中间有一小段静音。它们是用相同的乐器演奏的吗?	–
	两段音频片段之间有短暂的停顿。第二段音频片段使用的乐器与第一段相同，还是不同?	–
	给你两段音频片段，中间有一小段静音间隔。它们是否使用相同的乐器?	–
	请仔细聆听这两段音频片段，它们之间隔着一小段静音。相同的乐器 , 还是不同?	–
	两段音频片段依次播放，中间有一段停顿。是否是同一种乐器演奏了两次 , 还是两种不同的乐器?	–
节奏	给出两个音频片段，中间有一小段静音。哪个片段速度更快?	速度更快的音频片段
	给出两段音频片段，中间隔着一小段静音。哪段片段移动得更慢?	较慢的音频片段
	给出两段声音片段，中间隔着一小段静音间隔。哪段片段速度更快?	较快的音频片段
	现在为您播放两段声音片段，中间由一小段静音间隔隔开。哪一部分的速度更慢?	速度较慢的音频片段
	请聆听一对音频片段，中间有一小段静音间隔将它们分隔开来。哪段音频听起来更慢?	速度较慢的音频片段
	给你两段音频片段，中间有一小段静音间隔。哪一段移动得更快?	较快的音频片段
	给出两段听觉录音，中间由一小段停顿隔开。哪一部分的节奏更快?	较快的音频片段
	为你提供了两个音频片段，中间隔着一段短暂的静音间隔。哪一部分的速度更快?	速度更快的音频片段
	这里有两段声音录音，中间隔着一小段静音片段。哪段音频片段播放得更慢?	速度较慢的音频片段
	你会听到两个听觉片段，中间隔着一段短暂停顿。哪个片段播放得更快?	更快的音频片段

统计数据。 该基准测试包含 11,825 个生成项目：旋律、和声和音色各 3,025 个，节奏 2,750 个。旋律、和声和音色还包含额外的捕捉试次，其中两个音频片段完全相同；本文不使用这些试次。本文使用的题目包括 10,000 个训练项目、1,000 个干净测试项和 2,000 个翻转/保持项目（测试集划分中每项任务各有 250 个）。在这

表8：在两种字母含义下，固定指令均会逐字附加到每种措辞之后（音频占位符之前）。交换顺序仅适用于相同/不同任务，占其训练项目的 51.4%；测试集始终采用原始顺序。

适用于	指令（逐字）	
旋律/和声/音色，原始顺序	“相同仅回答‘A’，不同回答‘B’。解释、标点或额外文本。”	不要添加任何
旋律/和声/音色，交换顺序	“相同仅回答‘B’，不同回答‘A’。解释、标点或额外文本。”	不要添加任何
节奏，两个数据集划分	“第一个只回答‘A’，第二个只回答‘B’。解释、标点或额外文本。”	不要添加任何

1,000 个干净测试项中，528 个的答案是“A”，472 个的答案是“B”。真实测试集还增加了 600 个干净测试项和 1,200 个翻转/保持项目（附录 Q）。

音色库重叠。 MuseScore General 源自 FluidR3Mono，即 Flu-idR3 的单声道版本。其文档说明，未单独注明来源的乐器仍使用 FluidR3Mono 采样，而其更新日志列出了新的钢琴采样 [5]。对于仍保留 FluidR3Mono 采样的乐器，测试音色库更改了文件及其乐器设置，但源录音与两个训练音色库之一 FluidR3 GM 存在重叠。因此，对于这些乐器而言，音色库的更改力度弱于钢琴。真实测试集（附录 Q）与任何音色库均不共享音频。

B 人类参考研究

共有12人参与，他们的音乐专业水平各不相同。每人回答24个测试项目，因此我们共收集到288个答案。他们听到的音频文件和问题与模型听到的相同。我们剔除了三轮测试结果和一份空白提交。在干净题目上的准确率为87.5%（96题中答对84题；95%区间为79.4至92.7），在翻转项上的准确率为94.8%（96题中答对91题；88.4至97.8），在保持项上的准确率为79.2%（96题中答对76题；70.0至86.1）。按任务统计，在全部题目中，音色任务的准确率为97.2%（72题中答对70题），节奏任务为94.4%（72题中答对68题），旋律任务为83.3%（72题中答对60题），和声任务为73.6%（72题中答对53题）。每位参与者的准确率介于66.7%至95.8%之间。区间是根据答案计算的威尔逊区间。由于每人回答了24题，我们还对听众进行了重抽样：以有放回的方式抽取12位听众，得到翻转/保持准确率的95%区间为82.3%至91.1%。测试面向所有人开放。我们没有根据音乐背景筛选听众，也没有控制参与者人选；每位听众从四个音乐经验等级中选择一个：休闲听众；业余音乐家（会演奏或演唱，正规训练较少）；受过训练的音乐家（接受课程学习五年或以上，或拥有音乐学位）；音乐或音频研究人员。四位听众是休闲听众，两位是业余音乐家，四位是受过训练的音乐家，两位是音乐研究人员。在翻转/保持项目上，休闲听众的得分为82.8%（53/64），有音乐经验的八位听众得分为89.1%（114/128）；在干净题目上，两组得分均为87.5%（分别为28/32和56/64）。12位听众听到的都是相同的24个项目，他们之间的一致性很高（所有项目的 Fleiss’ κ 0.71，翻转/保持项目为0.76）。

表 9：按任务统计的人类回答。干净：每项任务 24 个回答。翻转/保持：每项任务 48 个回答，包含两个变体。所有任务合计有 5 次漏判和 20 次错误翻转，因此错误翻转中至少有 11 次发生在和声任务上，而音色和节奏任务各自至多有 1 次。

	干净		翻转/保持	
Task	正确	%	正确	%
旋律	24 个中的 19 个	79.2	48 个中的 41 个	85.4
和声	24 个中的 21 个	87.5	48 个中的 32 个	66.7
音色	24 个中的 23 个	95.8	48 个中的 47 个	97.9
节奏	24 个中的 21 个	87.5	48 个中的 47 个	97.9
All	96 个中的 84 个	87.5	192 个中的 167 个	87.0

[5] https://github.com/CarlGao4/Muse-Sounds

C 模型与推理

开源模型。 所有开源模型均在未经训练的情况下，使用各自默认的聊天模板进行测试。每个项目由一个音频文件和一个文本问题组成。我们比较答案首个位置上词元‘A’和‘B’的对数概率，并选择概率较大的那个。贪心解码在每个项目上都生成相同的字母。表10列出了八个模型各自的确切检查点（Hugging Face 名称和修订版本）、参数数量及音频采样率。

表10：开源模型检查点。修订版本为实际用于评估的本地 Hugging Face 缓存提交；参数数量根据检查点的 safetensors 文件头计算。

模型	Hugging Face 检查点（修订版本）	参数	音频采样率
Qwen2-Audio	Qwen/Qwen2-Audio-7B-Instruct (0a09522)	8.4B	16 千赫
Qwen2.5-Omni	Qwen/Qwen2.5-Omni-7B (ae9e169)	10.7B†	16 千赫
Audio Flamingo 3	nvidia/audio-flamingo-3-hf (7d4bae6)	16.5B	16 千赫‡
Phi-4-MM	microsoft/Phi-4-multimodal-instruct (93f923e)	5.6B	16 千赫
Kimi-Audio	moonshotai/Kimi-Audio-7B-Instruct (9a82a84)	∼7.9B§	16 千赫
MiMo-Audio	XiaomiMiMo/MiMo-Audio-7B-Instruct (c359441)	8.0B¶	24 千赫
Fun-Audio-Chat	FunAudioLLM/Fun-Audio-Chat-8B (7bf72dc)	9.5B	16 千赫
Step-Audio 2 mini	stepfun-ai/Step-Audio-2-mini (e36fdd5)	8.3B‖	16 千赫

†完整检查点（thinker 8.9B + talker 1.4B + token2wav 0.5B）；评估时仅实例化 thinker。 ‡据模型说明卡；缓存的快照中没有处理器配置可供确认。 §模型说明卡中的数字；尚未完成本地计数。 ¶不包括单独的 MiMo-Audio-Tokenizer（1.3B，修订版 5df9914）。 ‖不包括 token2wav 声码器权重。

商业模型。GPT-Audio-1.5 通过 OpenAI API 调用，使用的模型为 gpt-audio-1.5。Gemini 3.8 Flash 通过 Gemini API 调用，使用的模型为 gemini-3.8-flash （高强度思考）。两者的温度均设为 0。每次试验中的两个音频片段会拼接为一个单声道、16 位、44.1 千赫的 WAV 文件，并与问题一同以内嵌（base64）形式发送。这两个模型的 API 均不提供选项的对数概率，因此都依据生成的答案字母评分。对于 Gemini，我们将解码限制为两个选项字母，因此每次回复都是有效字母；没有出现无效答案（3,000 次调用中为 0 次）。Chat Completions 不支持对音频模型施加此类限制，因此 GPT-Audio-1.5 最多返回 16 个词元的补全内容，我们从中读取第一个独立出现的选项字母（以词边界判定，不区分大小写）。不包含独立字母的回复（干净条件下占 10.3%，翻转/保持条件下占 6.9%）记录为预测第一个选项；在答案键均衡的情况下，这种回退方式的得分处于随机水平，所有此类题目均已在发布的日志中标记。GPT-Audio-mini 和 Gemini 2.5 Pro 通过 OpenRouter API（OpenRouter 团队，2026年）调用，使用的模型 ID 分别为 openai/gpt-audio-mini 和 google/gemini-2.5-pro，所用 WAV 文件和提示词相同。我们从生成文本中读取答案字母。不包含答案字母的回复记录为第一个选项（“A”），与 GPT-Audio-1.5 的处理方式相同。GPT-Audio-mini 的 3,000 条回复中有 1,385 条此类回复，Gemini 2.5 Pro 则有 21 条。附录 M 也将这些回复计为错误。所有 API 运行均于 2026年9月进行。

D GRPO 详情

GRPO 仅使用训练集划分中的干净基底配对数据；表 5 中的仅声音版本均不用于训练。唯一的例外是附录 K 中的对照实验 R1，该实验将这些版本纳入其中。除 Audio Flamingo 3 的全量微调使用第500步的检查点外，我们报告所有运行在第 1,000 步的检查点。留一任务运行使用 LoRA 在 7,500 个题目上训练 500 步，并采用最终检查点。

表11：GRPO 训练详情。批次大小以每步每个图形处理器的提示词数（B）和每个提示词的滚动采样数（组大小 G）表示；所有运行每步都进行一次梯度更新。实际耗时是在共享节点上测得的耗时（参见 脚注11）。Q2A = Qwen2-Audio-7B-Instruct，Omni =Qwen2.5-Omni-7B，AF3 =Audio-Flamingo-3，Phi-4 = Phi-4-Multimodal。

	LoRA	全量微调
组 size G（滚动采样-数/提示词）	8	8
采样温度 / top-p	1.0 / 1.0	1。0 / 1。0
批次 size B（提示词/图形处理器/步）	4	4
有效批次（滚动输出/步）	64（8×8）[a]	64（8×8）
学习率（调度）	1×10^{-5}，常量	1×10^{-6}，常量
优化器	AdamW	Adafactor
梯度裁剪	1.0	1.0
KL 系数（相对于冻结参考模型）	0.10	0.10
LoRA 秩 / α / 丢弃率	16 / 32 / 0	—
可训练参数	仅语言模型解码器线性投影[b]	语言模型解码器（+语言模型头）
最大答案长度	8 个新生成词元	8 个新生成词元
训练步数	1000（全部模型）	1000（Q2A、Omni、Phi-4）；500（AF3）
硬件	每次运行使用 2 个图形处理器[c]	2× NVIDIA A100（80 GB）
实际耗时（小时）	Q2A ≈2.3、Omni ≈5.9、AF3 ≈3.9，Phi-4 ≈0.8	Q2A ≈2.7、Omni ≈2.8、AF3 ≈2.7，Phi-4 ≈25.5[d]

奖励（每次滚动采样）：$r = 1.0 \cdot \mathbb{1}[\text{正确}] + 0.2 \cdot \mathbb{1}[\text{解析为 A/B}] + 0.1 \cdot s_{\text{task}} \cdot$（$\pm1$），其中最后一项是反坍缩 nudging（$+$ 如果该次滚动采样预测了该任务当前代表性不足的字母，$-$ 否则；s_{task} 是每个任务预测字母偏斜度的指数移动平均，衰减率为 0。95）。零优势组最多会被重新采样 4 次，之后丢弃。

[a]Qwen2.5-Omni LoRA 运行于 6 个图形处理器上：每步 24 个提示词，$\times$ 8 = 192 次滚动采样。b语言模型解码器的注意力和 MLP 投影层（$q/k/v/o$/gate/up/down；Phi-4 中为融合的 qkv/o/gate up/down）；音频编码器和投影器冻结。c单个 8-GPU 节点：2$\times$ A100（80 GB）$+$ 6$\times$ A40（46GB）；每次 LoRA 运行中 A100/A40 的确切分配未记录。d节点与其他任务共享；从首次到最后一次检查点写入之间的经过时间；Phi-4 全量微调运行与多个并发任务重叠，因此其经过时间相较于独占 GPU 计算明显偏长（参见其他全量微调运行的 ≈2.7–2.8 小时）。

E 排行榜提交格式

每份提交都是一个 JSON Lines 文件，每个测试条目对应一条记录，其中包含条目 ID 和答案字母（‘A’或‘B’）。记录还需包含模型名称和版本、答案的读取方式（字母概率或生成文本）、模型是否以任何形式使用了我们的训练集划分、运行日期，以及可选的代码链接。缺失或无效的答案均计为错误。两个测试集使用相同的格式。评分脚本与表 2 和表 3 所用的相同。不允许在测试条目上进行训练。

F 完整结果

图4展示了每个模型在翻转/保持两种错误率上的位置，图5展示了四个经过训练的模型的字母偏差。下表给出了表 2 和表 3 背后的各任务具体数据。

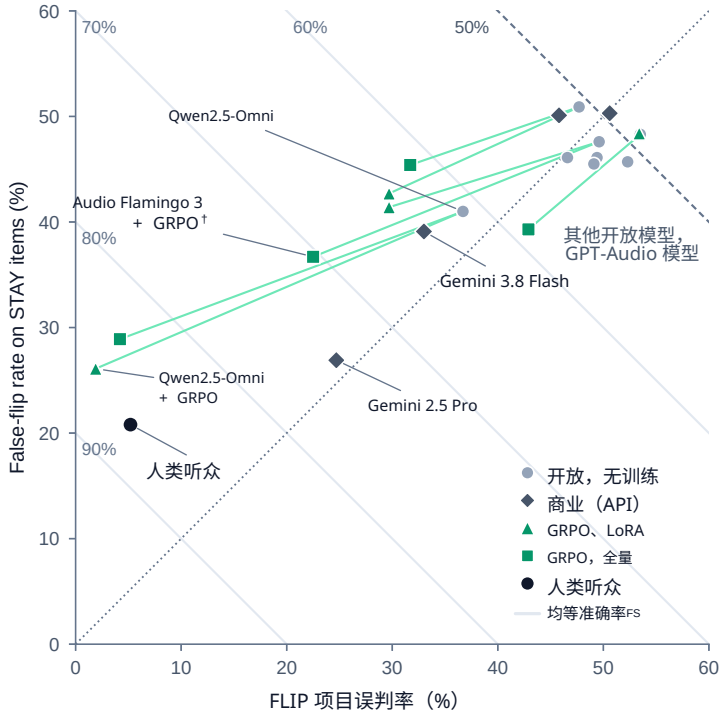

图4：GRPO 前后所有模型在翻转/保持两种错误率上的表现。沿着每条灰色对角线，准确率 Acc_{FS}保持不变；忽略音频的模型位于虚线对角线（50%）上。灰色圆圈：未经训练的开放模型。深色菱形：商业模型。绿色：GRPO 运行结果（三角形表示 LoRA，正方形表示全量微调），并以细线连接至其未经训练的模型。黑色：我们的听众。未经训练时，能够听辨音频的模型（Gemini 2.5 Pro、Gemini 3.8 Flash、Qwen2.5-Omni）位于相等比率的点线附近。GRPO 会使模型向左移动，减少漏判；相比之下，向下移动的幅度小得多：最佳运行结果将漏判率降至 1.9%，但错误翻转率仍为 26.1%，与 Gemini 2.5 Pro（26.9%）相当。 †第500步。

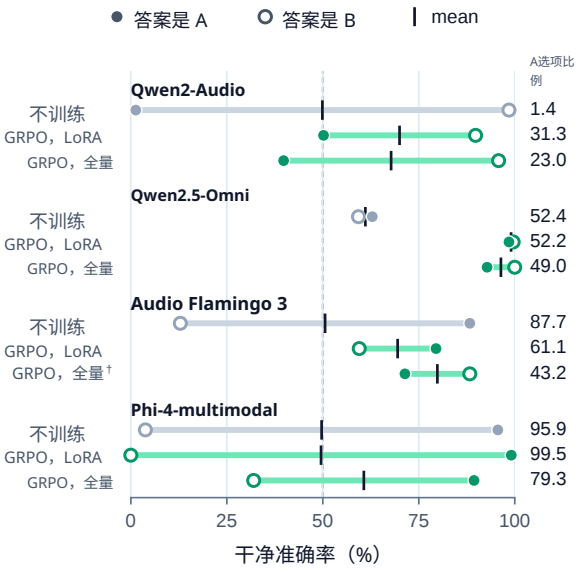

图5：干净测试集上的字母偏差。实心点：正确答案为“A”的题目的准确率。空心点：答案为“B”的题目的准确率。刻度线：二者的平均值，即类别平衡准确率。灰色行：无训练；绿色行：经过 GRPO 后。右列：“A”答案所占比例（正确答案中有 52.8% 为“A”）。
†第500步。未经训练时，四个模型中有三个位于音阶的两个端点附近，平均值接近 50%。对于采用全量微调的 Qwen2-Audio、Audio Flamingo 3 和 Phi-4-多模态模型，GRPO 使两个点更加接近；同时，Qwen2.5-Omni 的两个点均提升至 92% 以上。

表12：无训练时的每任务漏检率和错误翻转率（%）。商业行：附录I中逐条目分析的两个模型。

	MR				FFR			
模型	Mel.	Har.	Tim.	Rhy.	Mel.	Har.	Tim.	Rhy.
Qwen2-Audio	46.0	50.4	46.0	48.4	54.0	49.6	52.8	47.2
Qwen2.5-Omni	38.0	36.0	24.8	48.0	42.0	45.6	33.6	42.8
Audio Flamingo 3	50.4	49.6	54.4	44.0	46.4	52.4	41.2	50.4
Phi-4-multimodal	54.0	51.6	56.8	51.6	46.8	50.4	43.2	52.8
Kimi-Audio	53.2	49.2	56.0	50.8	41.2	52.4	42.8	46.4
MiMo-Audio	54.0	47.2	54.0	42.4	46.0	50.0	41.6	46.8
Fun-Audio-Chat	50.0	41.2	43.2	52.0	50.4	48.4	36.4	49.2
Step-Audio 2	55.2	49.2	42.0	50.0	49.2	46.4	38.0	48.4
GPT-Audio-mini	50.4	47.6	52.8	51.6	53.6	51.2	43.6	52.8
Gemini 2.5 Pro	23.2	35.2	19.2	21.2	30.4	37.2	16.0	24.0

表13：GRPO 后每任务漏检率和错误翻转率（%）。†第500步。

		MR				FFR			
模型	GRPO	Mel.	Har.	Tim.	Rhy.	Mel.	Har.	Tim.	Rhy.
Qwen2-Audio	LoRA	26.0	40.4	4.4	48.0	45.2	46.8	34.0	44.8
	full	29.6	41.6	6.8	48.8	46.8	45.2	40.8	48.8
Qwen2.5-Omni	LoRA	0.4	4.0	3.2	0.0	32.4	34.4	20.8	16.8
	full	0.0	15.2	1.2	0.4	34.0	41.2	22.8	17.6
Audio Flamingo 3	LoRA	40.4	24.4	2.4	51.6	50.8	41.2	20.4	53.2
	全量†	34.4	20.0	2.4	33.2	48.4	38.8	20.4	39.2
Phi-4-multimodal	LoRA	55.6	49.6	56.8	51.6	47.2	50.4	43.2	52.8
	full	48.8	45.6	25.6	51.6	48.4	46.0	10.0	52.8

表14：干净测试集上的字母偏差（%）。A选项比例：回答为“A”的题目占比（正确答案中有 52.8% 是“A”）。Acc|A 和 Acc|B：正确答案为“A”或“B”的题目的准确率。平衡值：两者的平均值。差值：两者的差异。最后两行是商业模型（附录I）。GPT-Audio-mini 的 A选项比例包含未提供答案字母的回复，这些回复被记录为“A”（附录 C）。

模型	GRPO	A选项比例	Acc|A	Acc|B	平衡值	Gap
Qwen2-Audio	none	1.4	1.3	98.5	49.9	97.2
	LoRA	31.3	50.2	89.8	70.0	39.6
	full	23.0	39.8	95.8	67.8	56.0
Qwen2.5-Omni	none	52.4	62.9	59.3	61.1	3.6
	LoRA	52.2	98.5	99.6	99.0	1.1
	full	49.0	92.8	100.0	96.4	7.2
Audio Flamingo 3	none	87.7	88.3	12.9	50.6	75.3
	LoRA	61.1	79.5	59.5	69.5	20.0
	full	43.2	71.4	88.3	79.8	17.0
Phi-4-multimodal	none	95.9	95.6	3.8	49.7	91.8
	LoRA	99.5	99.1	0.0	49.5	99.1
	full	79.3	89.4	32.0	60.7	57.4
GPT-Audio-mini	none	81.1	80.5	18.2	49.4	62.3
Gemini 2.5 Pro	none	60.8	87.9	69.5	78.7	18.4

表15：未经过训练时各任务的翻转/保持准确率（%）。模型：每项任务500道题，全部共2,000道；在单项任务中，得分介于45.6%至54.4%之间，与50%没有显著差异。人类听众：每项任务48个答案，全部共192个。

模型	Mel.	Har.	Tim.	Rhy.	All
Qwen2-Audio	50.0	50.0	50.6	52.2	50.7
Qwen2.5-Omni	60.0	59.2	70.8	54.6	61.2
Audio Flamingo 3	51.6	49.0	52.2	52.8	51.4
Phi-4-multimodal	49.6	49.0	50.0	47.8	49.1
Kimi-Audio	52.8	49.2	50.6	51.4	51.0
MiMo-Audio	50.0	51.4	52.2	55.4	52.2
Fun-Audio-Chat	49.8	55.2	60.2	49.4	53.6
Step-Audio 2	47.8	52.2	60.0	50.8	52.7
GPT-Audio-mini	48.0	50.6	51.8	47.8	49.6
GPT-Audio-1.5	50.4	51.4	56.0	50.4	52.1
Gemini 2.5 Pro	73.2	63.8	82.4	77.4	74.2
Gemini 3.8 Flash	57.2	51.2	84.2	63.2	64.0
人类 (12)	85.4	66.7	97.9	97.9	87.0

表16：从未训练到 GRPO 后漏检率和错误翻转率的下降幅度，以百分点计。 z 和 p（双侧）检验漏检率的降幅是否大于错误翻转率。该检验将训练前后的比率视为各包含 1,000 个题目的独立样本，因此在两者正相关时会高估方差。 †第500步。

模型	GRPO	ΔMR	ΔFFR	比率	z	p
Qwen2-Audio	LoRA	18.0	8.2	2.2	3.2	0.0015
	full	16.0	5.5	2.9	3.4	<0.001
Qwen2.5-Omni	LoRA	34.8	14.9	2.3	7.6	<0.001
	full	32.5	12.1	2.7	7.6	<0.001
Audio Flamingo 3	LoRA	19.9	6.2	3.2	4.4	<0.001
	全量†	27.1	10.9	2.5	5.4	<0.001
Phi-4-multimodal	LoRA	0.1	−0.1	n/a	0.1	0.95
	full	10.6	9.0	1.2	0.5	0.61

G 字幕生成详情

表17：在 2,735 个音频片段上，GRPO 前后的 MusicCaps 字幕生成（分数 $\times 100$）。MET.：METEOR；R-L：ROUGE-L；BERTS.：BERTScore-F1；Words：字幕平均长度。

模型	GRPO	BLEU-4	MET.	R-L	BERTS.	词语
Qwen2-Audio	none	13.06	32.09	33.43	39.27	40.7
	LoRA	13.38	32.42	33.74	39.73	40.8
	full	14.79	34.22	35.28	41.22	41.3
Qwen2.5-Omni	none	1.31	16.33	20.96	23.35	29.2
	LoRA	1.09	15.65	20.25	22.07	28.4
	full	1.60	16.69	21.23	23.50	30.0

我们使用 MusicCaps 中 is audioset_eval set 的数据行（2,858 个音频片段）。其中，2,735 个音频片段可供下载，并由每个模型版本生成了描述，因此所有比较都使用相同的音频片段。音频片段时长为 10 秒，采样率为 16 千赫，采用单声道。解码采用贪心策略，最多生成 256 个新生成词元，重复惩罚为 1.0。提示词为：“请用几句话描述这段音乐。请提及流派、乐器、情绪和节奏。”每个音频片段对应一条参考描述。没有任何模型版本生成退化描述。

表 18：在 2,735 个 MusicCaps 音频片段上，相对于基底模型的配对差异（Δ = 训练集 −基底，单位为分数点 $\times 100$）；方括号内为基于音频片段计算的 95% 自助法置信区间（10,000 次重采样）。星号表示 Wilcoxon 符号秩检验结果，并在每个模型的各项指标内进行经霍尔姆校正（$^{*}p < .05$，$^{**}p < .01$，$^{***}p < .001$）。BLEU-4† 为语料级指标，因此不适用逐音频片段的 Wilcoxon 检验。BERTScore-F1 使用 roberta-large（lang=en，默认第 17 层，使用 baseline=True 进行重缩放）。检查点：GRPO 第 1,000 步。

模型	变体	BLEU-4†	sBLEU	METEOR	ROUGE-L	BERTScore-F1
Qwen2-Audio	LoRA	$+0.3\,[+0.2, +0.5]$	$+0.3\,[+0.2, +0.5]^{**}$	$+0.3\,[+0.1, +0.5]^{*}$	$+0.3\,[+0.1, +0.5]^{***}$	$+0.5\,[+0.3, +0.6]^{***}$
	全量微调	$+1.7\,[+1.5, +2.0]$	$+1.8\,[+1.5, +2.1]^{***}$	$+2.1\,[+1.8, +2.5]^{***}$	$+1.8\,[+1.6, +2.1]^{***}$	$+2.0\,[+1.7, +2.3]^{***}$
Qwen2.5-Omni	LoRA	$-0.2\,[-0.3, -0.1]$	$-0.1\,[-0.2, -0.1]^{***}$	$-0.7\,[-0.8, -0.5]^{***}$	$-0.7\,[-0.8, -0.6]^{***}$	$-1.3\,[-1.5, -1.0]^{***}$
	全量微调	$+0.3\,[+0.2, +0.4]$	$+0.1\,[+0.1, +0.2]^{***}$	$+0.4\,[+0.2, +0.5]^{***}$	$+0.3\,[+0.1, +0.4]^{***}$	$+0.2\,[-0.1, +0.4]$

H 按声音变化统计的错误翻转

每组保持不变的配对都会将片段 B 替换为一个在某一方面有所不同的版本（第 3.3 节）：均衡器变化（± 3 dB 架），房间回声（混响时间为 0.2、0.5 或 1.0 秒）、背景嘶声（-50 dBFS），或者仅对旋律和和声进行最多 200 音分的移调。这 1,000 组保持不变的配对包含 294 组均衡器配对、290 组房间回声配对、292 组嘶声配对和 124 组移调配对。表 19 给出了四项任务合并后的各类变化错误翻转率，表 20 则按任务分别列出错误翻转率。两张表采用的答案与表 2 和表 3 相同。图 6 展示了 GRPO 如何改变合并后的比率。

接近随机水平的模型没有表现出任何规律。对于未经训练的开放模型（Qwen2.5-Omni 除外）以及两个 GPT-Audio 模型，无论变化类型如何，合并后的比率都在 42.8% 到 54.0% 之间。这些模型几乎无法区分这些音频片段，因此变化类型影响甚微。

最擅长忽略均衡器和房间回声的是能够听音的模型。三个未经训练、能够听音的模型在嘶声和移调配对上的错误率最高（32.5%至50.0%）。Qwen2.5-Omni 和 Gemini 3.8 Flash 在均衡器配对上的错误率最低（分别为33.7%和29.9%），而 Gemini 2.5 Pro 在房间回声配对上的错误率最低（17.2%，相比之下其在均衡器配对上为22.4%）。GRPO 保持了这一顺序，并扩大了差值。在每次学会相关任务的实验运行中，均衡器错误率下降了10.9至32.3个百分点，而嘶声错误率变化了 -6.2至 $+5.8$ 个百分点，移调错误率变化了 -4.0至 $+2.4$个百分点。

嘶声。经过 GRPO 后，四次 Qwen 实验运行在51.0%至53.1%的嘶声配对上出错，与训练前大致相同。两次 Qwen2.5-Omni 实验运行在63个旋律嘶声配对中的恰好38个，以及63个和声嘶声配对中的恰好30个上出错。这些是正确答案为“相同”的嘶声配对数量（其余为“不同”），

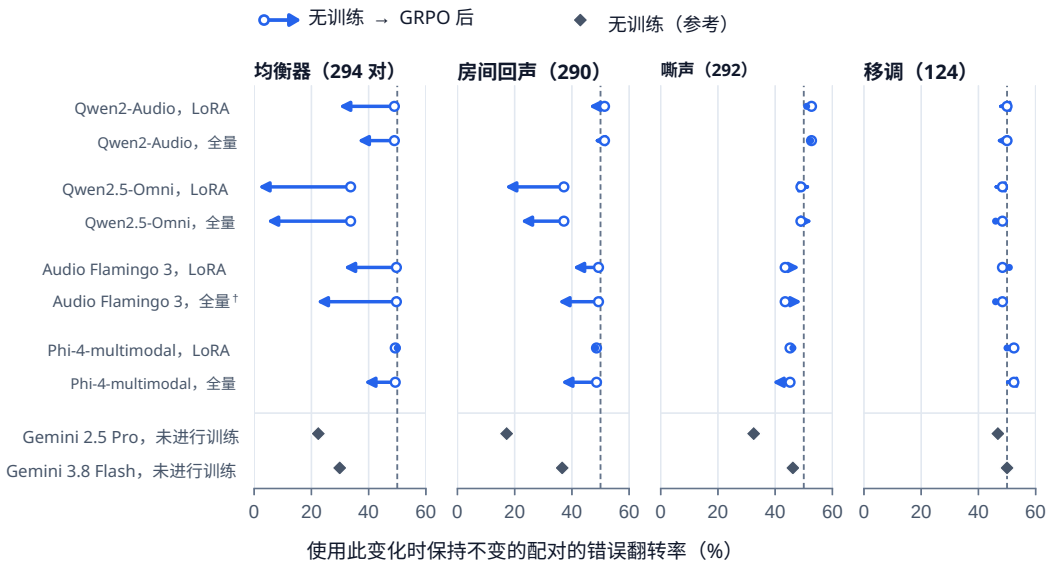

图 6：GRPO 前后按保持变化类型划分的错误翻转率，每次运行各占一行。每支蓝色箭头从未训练时的错误翻转率（空心圆）开始，指向 GRPO 后的错误翻转率；幅度较小的变化以线段表示，末端为圆点。菱形：未经训练的两个 Gemini模型。虚线：50%。在所有学会的运行中，GRPO 都降低了均衡器条件下的错误翻转率（降低 10.9 至 32.3 个百分点）；在房间回声条件下也有一定降低，但在嘶声或移调条件下没有降低。Gemini 2.5 Pro 未经任何训练也呈现相同的排序：房间回声和均衡器条件下的错误翻转率最低，嘶声条件下较高，移调条件下接近 50%。 †第500步。

表 19：按保持变化类型划分的错误翻转率（%），汇总所有任务；越低越好。房间：房间回声。括号内为配对数量。全部：全部 1,000 组保持不变的配对上的错误翻转率，与表 2 和表 3 相同。移调仅用于旋律和和声。 †第500步。

模型	GRPO	EQ (294)	Room (290)	Hiss (292)	透明度 (124)	All (1,000)
Qwen2-Audio	none	49.0	51.4	52.7	50.0	50.9
	LoRA	29.6	45.9	51.0	46.8	42.7
	full	36.1	47.6	52.4	46.0	45.4
Qwen2.5-Omni	none	33.7	37.2	49.0	48.4	41.0
	LoRA	1.4	16.6	52.4	45.2	26.1
	full	4.4	22.1	53.1	46.0	28.9
Audio Flamingo 3	none	49.7	49.3	43.5	48.4	47.6
	LoRA	31.3	40.3	48.6	50.8	41.4
	全量†	21.8	35.2	49.3	46.0	36.7
Phi-4-multimodal	none	49.3	48.6	45.2	52.4	48.3
	LoRA	50.0	48.3	46.2	50.0	48.4
	full	38.4	36.2	39.0	49.2	39.3
其他开放模型，未经训练						
Kimi-Audio	none	45.6	47.9	42.8	47.6	45.7
MiMo-Audio	none	47.6	43.1	46.6	48.4	46.1
Fun-Audio-Chat	none	45.9	45.9	45.2	49.2	46.1
Step-Audio 2	none	43.5	46.9	44.5	49.2	45.5
商业模型（*API*），未经训练						
GPT-Audio-mini	none	49.7	50.7	49.0	54.0	50.3
GPT-Audio-1.5	none	50.0	47.6	53.8	47.6	50.1
Gemini 2.5 Pro	none	22.4	17.2	32.5	46.8	26.9
Gemini 3.8 Flash	none	29.9	36.6	46.2	50.0	39.1

表 20：按任务和保持变化划分的错误翻转率（%）；越低越好。房间：房间回声；Tr.：移调。旋律和和声：63 个均衡器、62 个房间回声、63 个嘶声以及 62 个移调配对。音色和节奏：84 个均衡器、83 个房间回声和 83 个嘶声配对，且没有移调。†第500步。

		旋律				和声				音色			节奏		
模型	GRPO	均衡器	房间	嘶声	Tr.	均衡器	房间	嘶声	Tr.	均衡器	房间	嘶声	均衡器	房间	嘶声
Qwen2-Audio	none	49.2	58.1	60.3	48.4	47.6	51.6	47.6	51.6	52.4	53.0	53.0	46.4	44.6	50.6
	LoRA	20.6	56.5	60.3	43.5	38.1	51.6	47.6	50.0	8.3	37.3	56.6	51.2	42.2	41.0
	full	25.4	58.1	60.3	43.5	36.5	48.4	47.6	48.4	25.0	41.0	56.6	54.8	45.8	45.8
Qwen2.5-Omni	none	34.9	32.3	54.0	46.8	33.3	45.2	54.0	50.0	22.6	34.9	43.4	44.0	37.3	47.0
	LoRA	0.0	29.0	60.3	40.3	4.8	35.5	47.6	50.0	1.2	8.4	53.0	0.0	1.2	49.4
	full	0.0	35.5	60.3	40.3	19.0	46.8	47.6	51.6	1.2	10.8	56.6	0.0	4.8	48.2
Audio Flamingo3	无	49.2	51.6	34.9	50.0	57.1	51.6	54.0	46.8	41.7	42.2	39.8	52.4	53.0	45.8
	LoRA	41.3	56.5	60.3	45.2	28.6	33.9	46.0	56.5	3.6	18.1	39.8	53.6	55.4	50.6
	全量†	38.1	51.6	60.3	43.5	17.5	43.5	46.0	48.4	2.4	12.0	47.0	32.1	39.8	45.8
Phi-4-multimodal	none	50.8	43.5	41.3	51.6	50.8	51.6	46.0	53.2	42.9	43.4	43.4	53.6	55.4	49.4
	LoRA	52.4	45.2	39.7	51.6	52.4	48.4	52.4	48.4	42.9	43.4	43.4	53.6	55.4	49.4
	full	46.0	43.5	54.0	50.0	50.8	41.9	42.9	48.4	8.3	7.2	14.5	53.6	55.4	49.4
其他开放模型，无训练															
Kimi-Audio	none	41.3	43.5	39.7	40.3	54.0	50.0	50.8	54.8	42.9	47.0	38.6	45.2	50.6	43.4
MiMo-Audio	none	50.8	41.9	39.7	51.6	52.4	53.2	49.2	45.2	39.3	42.2	43.4	50.0	37.3	53.0
Fun-Audio-Chat	none	52.4	50.0	47.6	51.6	49.2	46.8	50.8	46.8	35.7	34.9	38.6	48.8	53.0	45.8
Step-Audio 2	none	55.6	45.2	47.6	48.4	42.9	43.5	49.2	50.0	35.7	41.0	37.3	42.9	56.6	45.8
商业模型（*API*），无训练															
GPT-Audio-mini	none	55.6	51.6	49.2	58.1	49.2	50.0	55.6	50.0	41.7	45.8	43.4	53.6	55.4	49.4
GPT-Audio-1.5	none	49.2	56.5	60.3	46.8	44.4	45.2	47.6	48.4	50.0	45.8	55.4	54.8	44.6	51.8
Gemini 2.5 Pro	none	22.2	17.7	34.9	46.8	42.9	25.8	33.3	46.8	13.1	14.5	20.5	16.7	13.3	42.2
Gemini 3.8 Flash	none	34.9	51.6	52.4	50.0	50.8	50.0	49.2	50.0	10.7	9.6	34.9	29.8	42.2	50.6

因此，这些计数表示：如果模型对这两个任务中的每一组嘶声配对都回答“不同”，无论音乐如何，它会给出的结果。在音色和节奏方面，嘶声导致了 Qwen2.5-Omni 使用 LoRA 时 94 次错误翻转中的 85 次，以及使用全量微调时 101 次中的 87 次。在这两个任务的均衡器和房间回声配对中，两次运行的错误率分别为 2.7% 和 4.2%。使用全量微调的 Phi-4-multimodal 是例外。在音色任务上——这是它唯一学会的任务——它在 14.5% 的嘶声配对上出错，而均衡器和房间回声的错误率分别为 8.3% 和 7.2%。这也是唯一一次漏判和错误翻转以相近幅度下降的运行（第5.3节）。

移调。没有模型学会忽略移调：无论是否经过训练，每次运行都会在 45.2% 至 54.0% 的移调配对上出错，包括 Gemini 2.5 Pro（46.8%）；它在房间回声配对上的错误率仅为 17.2%。经过 GRPO 训练后，Qwen2.5-Omni 在 99.6% （ LoRA ）和 100.0% （全量）的旋律翻转配对中检测出单个移动音符，但在两次运行中，对移调后的旋律对的错误率均为 40.3%。

房间回声取决于任务。经过 GRPO 训练后，使用 LoRA 的 Qwen2.5-Omni 在旋律和和声的房间回声配对上，错误率分别为 29.0% 和 35.5%；而在音色和节奏上的错误率分别为 8.4% 和 1.2%。

嘶声强度。在测试集中，均衡器和嘶声各自仅以一种强度应用。附录L中的嘶声扫描改变了嘶声强度：训练后的 Qwen2.5-Omni 已经会将 42.9% 的“相同”配对判为不同，强度为-80 dBFS。

I Gemini 2.5Pro 和 GPT-Audio-mini 的逐题分析

对于这两个商业模型，我们分析了每个答案，因此可以按正确答案和差异大小对其错误进行分类（表21和图7）。这些题目与表2中的1,000道干净题目和2,000道翻转/保持题目相同。

GPT-Audio-mini 大多回答“A”，或者不提供答案字母。它对81.1%的干净题目回答“A”，对全部500道节奏翻转/保持题目也都回答“A”，因此其节奏翻转/保持准确率为47.8%，即始终回答“A”所能取得的分数（第3.3 节）。其中许多 “A” 回答都没有附上答案字母（ 1,385

表21：旋律、和声和音色的错误率（%），按正确答案划分。相同：答案为“相同”的配对，若回答为“不同”则计为错误。不同。：答案为“不同”的配对，若回答为“相同”则计为错误。 n：配对数量。

	n		Gemini 2.5 Pro		GPT-Audio-mini	
题目	Same	不同。	Same	不同。	Same	不同。
干净	401	349	3.5	41.0	25.7	75.4
FLIP	349	401	4.3	44.6	24.1	73.1
保持，均衡器	109	101	6.4	44.6	22.0	76.2
STAY, 房间回声	115	92	2.6	39.1	28.7	73.9
STAY, hiss	115	94	10.4	51.1	27.8	74.5
保持、移调	62	62	51.6	41.9	41.9	66.1

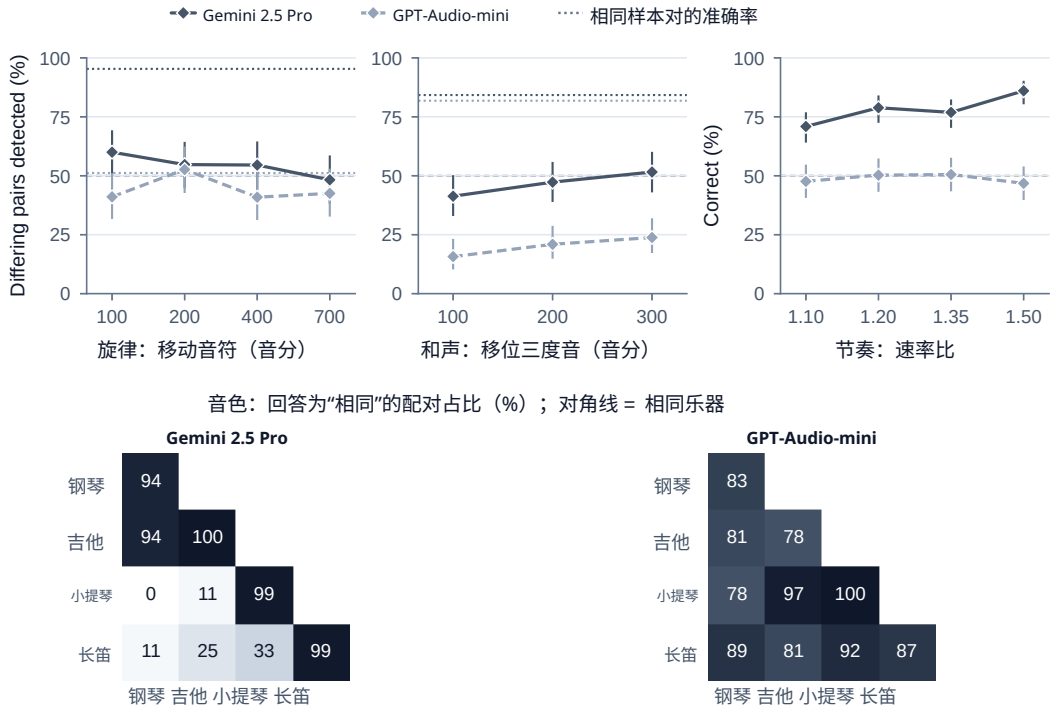

图7：Gemini 2.5 Pro 和 GPT-Audio-mini 对每个条目逐项分析，并汇总该条目的干净、翻转和保持版本。**上图：**模型将不同旋律对和和声对判为不同的比例，按变化幅度划分（虚线点：相同配对的准确率）；以及节奏准确率，按两条节拍音轨的速率比划分。每个数据点的柱形表示基于 87 至 189 个条目的 95% 威尔逊区间，虚线为 50%。**下图：**各乐器对中，音色配对被回答为“相同”的比例；对角线单元格为相同乐器配对，此时回答“相同”才是正确的。无论移动幅度如何，Gemini 2.5 Pro 对移动音符的检测率都约为一半；节奏差异越大，其表现越好；它还会将钢琴和吉他听成同一种乐器。GPT-Audio-mini 对大多数音色配对都回答“相同”，在旋律和节奏方面的表现则徘徊在 50% 左右。

在其 3,000 条回复中），表2将其记录为「A」（附录 C）。在音色方面，无论乐器对中的乐器是否不同，它对每一对都在 78% 至 100% 的情况下回答「相同」。在旋律和节奏方面，其准确率不会随差异大小而变化。

Gemini 2.5 Pro 倾向于回答「相同」。在三个相同/不同任务中，其干净错误有 91%、翻转/保持错误有 83% 是对不同乐器对回答「相同」。对于相同的乐器对，它几乎总能答对，但移调后例外（表21）。随着移动音符的距离增大，它对移动音符的检测并未改善：移动 100 音分的旋律对中，它检测出 60.0%；移动 700 音分的旋律对中，则检测出 48.3%。在和声方面，随着三度音移动距离增大，检测率从 41.3% 上升至 51.6%。只有在节奏任务中（这是唯一一个不涉及相同/不同判断的任务），准确率才会随着差异增大而明显上升，从速率比为 1.10 时的 70.9% 升至 1.50 时的 86.0%。在音色方面，它在 36 对样本中仅有 2 对能区分钢琴和吉他；但在全部 36 对中都能区分钢琴和小提琴，在 32 对中能区分钢琴和长笛。

声音变化很少会改变它的答案，移调以及节奏中的嘶声除外。在相同/不同任务中，它的大多数错误翻转都是对实际存在差异的保持不变的配对回答“相同”（占此类配对的39.1%至51.1%）；在干净和变化配对中，它也会犯同样的错误（分别为41.0%和44.6%）。对于实际相同的保持不变的配对，均衡器、房间回声和嘶声仅分别使它在6.4%、2.6%和10.4%的情况下回答“不同”，但移调会使这一比例达到51.6%。因此，在这些任务中，Gemini 2.5 Pro基本忽略了三种音效，但会将移调后的旋律或和弦视为不同的旋律或和弦。在节奏任务中，嘶声将其准确率降至57.8%，而使用均衡器和房间回声时分别为83.3%和86.7%；其错误方向相反（83个嘶声配对中有17个和18个分别属于两种错误）。因此，它的错误翻转率与训练后的 Qwen2.5-Omni相近（分别为26.9%和26.1%），但错误类型不同：Gemini 2.5 Pro主要漏判实际存在的差异，而训练后的模型会对嘶声产生反应。

J Slakh2100 测试

构建。 本测试旨在考察：当真实歌曲的声音更加丰富、但标签仍然准确时，我们的发现是否依然成立。Slakh2100（Manilow 等，2019）使用商业采样乐器，将 MIDI 渲染成歌曲，因此每条音轨的 MIDI 都能提供所有标签。我们按照渲染集的方式构建测试集（附录 A）。任务、片段时长、文件格式、问题措辞和字母规则均相同。每项任务包含 250 个来自 Slakh 测试曲目的干净测试项，每个测试项各有一个翻转版本和一个保持版本。保持变化遵循表 6，并使用我们的生成器代码。训练集划分中，每项任务包含 2,500 个来自 Slakh 训练曲目和验证曲目的测试项。训练项与测试项之间没有共享曲目；旋律项也不会共享乐器插件。

- **旋律**。 从一条每次只演奏一个音符的音轨中（通过 MIDI 核对）截取一段由 4 到 6 个音符组成的乐句，时长为 2.5 秒。在“不同”配对中，使用 Rubber Band 将第二个音符移调 100、200、400 或 700 音分，同时保留共振峰。在“相同”配对中，同一个音符经过同一移调器处理，移调幅度为 0 音分，因此两种配对都经过相同的处理。
- **和声**。 从钢琴、吉他、管风琴或弦乐音轨中录制的单个音符。我们使用相同的移调器，以此构建三和弦，时长为 2.0 秒。在“不同”配对中，三度音会移动 100、200 或 300 音分，与我们的生成器设置相同。
- **音色**。 选取钢琴、原声吉他、小提琴、弦乐合奏或长笛音轨中的乐段，时长为 2.5 秒。两个音频片段始终是不同的乐段。“相同”配对取自同一音轨的两个乐段，“不同”配对则取自两个不同乐器家族的音轨。
- **节奏**。 鼓轨经 1.10、1.20、1.35 或 1.50 倍时间拉伸后，时长为 4.0 秒。较慢的音频片段也会经过拉伸器处理，倍率为 1.0。两个音频片段来自歌曲的同一部分，翻转会交换它们的位置。

评分。 评分方式与渲染测试集相同，但有一点不同：在这里，没有给出答案字母的回复计为错误，而表2将其记为“A”（附录 C）。在渲染测试中，这几乎没有影响：Gemini 2.5 Pro 在翻转/保持项目上给出了 16 条此类回复，其准确率因此为 73.7%，而非 74.2%。在 Slakh2100 翻转/保持项目上，此类回复更常见：GPT-Audio-mini、GPT-Audio-1.5 和 Gemini 2.5 Pro 的比例分别为 54.6%、8.4% 和 5.5%（在渲染测试中分别为 45.2%、6.9% 和 0.8%）。对于它们确实作答的题目，这些模型的得分分别为 51.5%、50.7% 和

79.7%。如果像表2那样将此类回复记为“A”，Gemini 2.5 Pro 的得分则为 78.2%。开放模型总是会给出一个字母。

结果。 在未进行训练的情况下（表22），结果与渲染测试一致。12个模型的排名几乎相同（斯皮尔曼 $\rho = 0.93$、$p < 0.001$；若将不含字母的回答计为“A”，则为0.86；在8个开放模型上则为0.76）。只有Qwen2.5-Omni（56.5%）、Fun-Audio-Chat（56.6%）和两个Gemini模型在翻转/保持任务上的表现明显高于随机水平（双侧二项检验，$p < 0.001$）。Audio Flamingo 3的得分为52.2%（$p = 0.05$），其他开放模型则为49.8%至51.4%。

表23加入了每种训练配置，包括在我们的渲染数据集划分上训练的检查点，以及在Slakh2100训练集划分上进行的新GRPO运行（设置参见表11，训练1,000步；Audio Flamingo 3采用全量微调时训练500步）。在我们的渲染配对上训练后，Qwen2.5-Omni（使用LoRA时为72.3%，采用全量微调时为70.9%）和Audio Flamingo 3（分别为59.0%和61.7%）的能力得以迁移。在Slakh2100本身上训练，则将Qwen2.5-Omni提升至86.0%（LoRA）和80.8%（全量）。对于两次运行，漏判再次比错误翻转下降得更多（表26）。Qwen2-Audio和Audio Flamingo 3的较弱运行结果仅达到56.2%至58.1%；对它们而言，错误翻转的下降幅度略高于漏判。与渲染测试一样，移调是训练模型最容易失败的保持变化。例如，在Slakh2100上使用LoRA训练的Qwen2.5-Omni，在均衡器变化下产生7.8%的错误翻转，而在移调下则为39.5%。表24和表25给出了各任务错误率和字母偏差。

表22：未经训练的 Slakh2100测试（%；该数据集上的表2和表15）。干净：每项任务250个题目（全部共1,000个）。翻转/保持：每项任务500个题目（$\mathrm{Acc}_{\mathrm{FS}}$共2,000个）。未包含答案字母的回复计为错误。

	干净准确率					翻转/保持准确率							
模型	Mel.	和声	Tim.	Rhy.	全部	旋律	和声	Tim.	Rhy.	$\mathrm{Acc}_{\mathrm{FS}}$	MR	FFR
Qwen2-Audio	50.0	50.4	50.4	50.0	50.2	50.0	50.0	50.8	54.0	51.2	48.0	49.6
Qwen2.5-Omni	54.4	50.8	66.0	52.0	55.8	55.0	51.2	63.6	56.0	56.5	43.7	43.4
Audio Flamingo 3	48.8	50.8	60.8	52.0	53.1	49.6	49.0	58.8	51.4	52.2	48.6	47.0
Phi-4-multimodal	50.0	48.4	50.8	50.0	49.8	51.2	51.8	50.0	46.0	49.8	51.2	49.3
Kimi-Audio	51.6	51.2	51.2	51.2	51.3	49.2	49.6	51.4	51.8	50.5	50.7	48.3
MiMo-Audio	50.0	51.6	52.8	57.6	53.0	50.0	53.2	52.4	50.0	51.4	48.5	48.7
Fun-Audio-Chat	54.4	55.2	60.0	54.0	55.9	52.2	53.6	63.8	56.8	56.6	43.3	43.5
Step-Audio 2	48.8	54.0	56.4	49.2	52.1	48.4	51.8	56.0	46.6	50.7	49.8	48.8
GPT-Audio-mini	38.0	29.6	19.6	5.2	23.1	37.8	31.2	18.4	6.0	23.4	76.7	76.6
GPT-Audio-1.5	47.6	42.0	52.0	44.0	46.4	48.8	43.8	50.6	42.4	46.4	53.9	53.3
Gemini 2.5 Pro	68.8	68.8	89.6	79.2	76.6	63.2	64.6	90.8	82.6	75.3	24.2	25.2
Gemini 3.8 Flash	58.0	54.0	84.0	63.6	64.9	52.8	51.2	83.2	59.0	61.6	38.1	38.8
始终选择“A”					50.2					49.0	51.9	50.1

表23：Slakh2100测试集上的所有开源模型配置（%；该数据集上的表3和表19）。Mel. 至全部：干净准确率。均衡器至 Tr.：每次保持变化的错误翻转率（Tr.：移调）。Slakh2100划分：10,000个训练项目，设置参见表11。R1：在干净的 + 仅声音配对上进行GRPO；R2：在干净配对上进行监督微调。†第500步；其他所有运行均为第1,000步。

模型	训练	旋律	和声	Tim.	Rhy.	全部准确率$_{FS}$	MR	FFR	均衡器	房间	Hiss	Tr.
无训练												
Qwen2-Audio	none	50.0	50.4	50.4	50.0	50.2	51.2	48.0	49.6	54.6	53.1	44.3	41.9
Qwen2.5-Omni	none	54.4	50.8	66.0	52.0	55.8	56.5	43.7	43.4	45.1	40.4	39.1	56.5
Audio Flamingo 3	无	48.8	50.8	60.8	52.0	53.1	52.2	48.6	47.0	42.4	43.2	49.8	60.5
Phi-4-multimodal	none	50.0	48.4	50.8	50.0	49.8	49.8	51.2	49.3	44.4	46.2	54.0	57.3
在渲染划分上进行 *GRPO*（表*3*）												
Qwen2-Audio	LoRA	50.4	55.2	52.0	48.0	51.4	52.6	46.3	48.5	52.9	52.7	42.6	41.9
	full	50.0	53.6	54.0	50.4	52.0	52.8	46.6	47.8	51.5	50.3	43.6	42.7
Qwen2.5-Omni	LoRA	80.0	86.4	64.4	72.0	75.7	72.3	25.3	30.2	24.1	30.8	32.2	38.7
	full	82.8	84.4	65.2	72.4	76.2	70.9	27.0	31.3	25.4	32.2	33.6	37.9
Audio Flamingo 3	LoRA	57.6	69.2	72.4	50.0	62.3	59.0	40.0	42.0	34.9	45.5	44.3	45.2
	全量†	60.0	64.4	80.0	53.2	64.4	61.7	36.4	40.3	34.6	43.5	41.2	44.4
Phi-4-multimodal	LoRA	49.6	50.0	50.8	50.0	50.1	49.1	52.0	49.8	43.4	46.6	54.7	61.3
	full	50.8	53.6	56.4	49.2	52.5	51.5	48.9	48.2	45.4	47.3	49.8	53.2
在 *Slakh2100*划分上使用 *GRPO*												
Qwen2-Audio	LoRA	50.8	58.4	72.8	49.6	57.9	56.8	43.7	42.8	43.4	47.3	38.8	40.3
	full	51.2	54.4	73.2	49.2	57.0	56.2	43.7	44.0	46.8	47.9	39.1	39.5
Qwen2.5-Omni	LoRA	89.2	88.8	96.4	95.6	92.5	86.0	10.4	17.6	7.8	18.2	17.6	39.5
	full	83.6	88.4	95.2	75.2	85.6	80.8	16.7	21.8	13.9	20.2	24.2	38.7
Audio Flamingo 3	LoRA	55.2	60.8	78.8	56.0	62.7	58.1	43.5	40.4	35.9	42.8	42.9	39.5
	全量†	54.0	54.8	78.4	50.0	59.3	57.2	45.8	39.9	32.5	38.0	43.6	53.2
Phi-4-multimodal	LoRA	49.6	50.0	50.8	50.0	50.1	49.1	52.0	49.8	43.4	46.6	54.7	61.3
	full	48.8	55.2	53.2	50.0	51.8	51.7	47.9	48.7	44.4	50.7	51.9	46.8
附录 K 的基线（合成配对）												
Qwen2.5-Omni	R1	74.0	79.6	96.4	76.4	81.6	77.1	22.1	23.7	17.3	18.8	26.0	45.2
	R2	84.8	86.8	64.0	72.8	77.1	70.9	26.9	31.3	25.1	32.5	33.6	37.9

表24：Slakh2100测试集上的每任务漏检率和错误翻转率（%；该测试集上的表12和表13）。每个区块中的列：旋律、和声、音色、节奏。未训练行：表12中的模型。†第500步。

		MR				FFR			
模型	GRPO	旋律	和声	音色	节奏	旋律	和声	音色	节奏
未训练									
Qwen2-Audio	none	50.0	50.0	48.8	43.2	50.0	50.0	49.6	48.8
Qwen2.5-Omni	none	44.0	50.0	40.4	40.4	46.0	47.6	32.4	47.6
Audio Flamingo3	无	48.4	51.6	44.0	50.4	52.4	50.4	38.4	46.8
Phi-4-multimodal	none	47.6	49.6	50.8	56.8	50.0	46.8	49.2	51.2
Kimi-Audio	none	52.8	49.6	49.6	50.8	48.8	51.2	47.6	45.6
MiMo-Audio	none	50.0	46.4	48.0	49.6	50.0	47.2	47.2	50.4
Fun-Audio-Chat	none	44.0	47.6	37.2	44.4	51.6	45.2	35.2	42.0
Step-Audio 2	none	49.2	49.2	45.2	55.6	54.0	47.2	42.8	51.2
GPT-Audio-mini	none	60.8	69.6	81.2	95.2	63.6	68.0	82.0	92.8
Gemini 2.5 Pro	none	34.4	36.8	8.0	17.6	39.2	34.0	10.4	17.2
在渲染划分上进行 *GRPO*									
Qwen2-Audio	LoRA	44.0	48.4	48.4	44.4	50.4	47.6	48.8	47.2
	full	48.4	46.4	46.4	45.2	50.0	47.6	47.6	46.0
Qwen2.5-Omni	LoRA	21.6	17.6	34.4	27.6	31.2	26.8	34.8	28.0
	full	22.4	20.4	35.6	29.6	28.0	32.0	35.6	29.6
Audio Flamingo3	LoRA	42.0	40.8	20.4	56.8	46.0	41.2	29.6	51.2
	全量†	46.8	33.2	15.2	50.4	46.0	43.6	24.8	46.8
Phi-4-multimodal	LoRA	50.8	49.2	51.2	56.8	48.4	49.6	50.0	51.2
	full	46.4	49.2	44.8	55.2	51.6	47.2	43.6	50.4
*Slakh2100*划分上的 *GRPO*									
Qwen2-Audio	LoRA	46.8	44.0	40.8	43.2	47.6	46.4	28.4	48.8
	full	48.8	44.4	38.8	42.8	50.0	46.8	30.4	48.8
Qwen2.5-Omni	LoRA	13.2	15.2	9.6	3.6	26.4	29.2	6.0	8.8
	full	18.0	16.0	11.6	21.2	28.8	28.8	6.4	23.2
Audio Flamingo 3	LoRA	47.2	38.4	37.6	50.8	49.6	42.4	21.2	48.4
	全量†	45.2	42.8	38.4	56.8	42.0	46.8	21.2	49.6
Phi-4-multimodal	LoRA	50.8	49.2	51.2	56.8	48.4	49.6	50.0	51.2
	full	43.2	41.6	50.0	56.8	49.2	46.4	48.0	51.2

表25：Slakh2100 干净测试集上的字母偏差（%；该数据集上的表14）。A选项比例：答案为“A”的题目占比（正确答案中有50.2%为“A”）。平衡值：Acc|A 和 Acc|B 的平均值。差值：Acc|A 与 Acc|B 之差的绝对值。 †第500步。

模型	GRPO	A选项比例	Acc|A	Acc|B	平衡值	差值
无，以及在渲染划分上的 *GRPO*						
Qwen2-Audio	none	0.6	1.0	99.8	50.4	98.8
	LoRA	6.0	7.6	95.6	51.6	88.0
	full	5.6	7.8	96.6	52.2	88.8
Qwen2.5-Omni	none	32.6	38.4	73.3	55.9	34.8
	LoRA	44.5	70.1	81.3	75.7	11.2
	full	40.2	66.3	86.1	76.2	19.8
Audio Flamingo 3	none	83.1	86.1	19.9	53.0	66.2
	LoRA	59.3	71.5	53.0	62.3	18.5
	全量†	55.0	69.3	59.4	64.4	9.9
Phi-4-multimodal	none	98.0	97.6	1.6	49.6	96.0
	LoRA	96.9	96.8	3.0	49.9	93.8
	full	87.3	89.6	15.1	52.4	74.6
GPT-Audio-mini	none	35.0	34.7	11.4	23.1	23.2
Gemini 2.5 Pro	none	57.2	86.7	66.5	76.6	20.2
在 *Slakh2100* 划分上使用 *GRPO*						
Qwen2-Audio	LoRA	12.1	20.1	96.0	58.1	75.9
	full	12.2	19.3	95.0	57.2	75.7
Qwen2.5-Omni	LoRA	48.5	90.8	94.2	92.5	3.3
	full	49.0	84.5	86.7	85.6	2.3
Audio Flamingo3	LoRA	42.1	54.8	70.7	62.7	15.9
	全量†	69.9	79.1	39.4	59.2	39.7
Phi-4-multimodal	LoRA	96.9	96.8	3.0	49.9	93.8
	full	84.6	86.3	17.1	51.7	69.2

表26：在 Slakh2100测试 上，从未训练到 GRPO 后的漏检率和错误翻转率下降幅度（以百分点计）（该数据集划分见表16）。 z 和 p：‘漏检率下降幅度大于错误翻转率’的双侧检验。 †第500步。

模型	GRPO	Δ漏检率	ΔFFR 比率		z	p
在渲染划分上进行 *GRPO*						
Qwen2-Audio	LoRA	1.7	1.1	1.5	0.2	0.85
	full	1.4	1.8	0.8	-0.1	0.90
Qwen2.5-Omni	LoRA	18.4	13.2	1.4	1.7	0.08
	full	16.7	12.1	1.4	1.5	0.13
Audio Flamingo 3	LoRA	8.6	5.0	1.7	1.1	0.25
	全量†	12.2	6.7	1.8	1.8	0.08
Phi-4-multimodal	LoRA	-0.8	-0.5	不适用	-0.1	0.92
	full	2.3	1.1	2.1	0.4	0.70
在 *Slakh2100*划分上使用 *GRPO*						
Qwen2-Audio	LoRA	4.3	6.8	0.6	-0.8	0.43
	full	4.3	5.6	0.8	-0.4	0.68
Qwen2.5-Omni	LoRA	33.3	25.8	1.3	2.8	0.0055
	full	27.0	21.6	1.2	1.9	0.06
Audio Flamingo 3	LoRA	5.1	6.6	0.8	-0.5	0.63
	全量†	2.8	7.1	0.4	-1.4	0.17
Phi-4-multimodal	LoRA	-0.8	-0.5	不适用	-0.1	0.92
	full	3.3	0.6	5.5	0.9	0.39

K 仅声音配对训练与监督微调

我们的主要实验仅使用干净配对进行训练（第5.3节）。本附录探讨两个问题。第5.3节中的模式——漏判大幅减少，远多于错误翻转的减少——是否依赖于 GRPO，还是任何使用干净配对进行的训练都会出现这种模式？加入仅声音变化的配对是否有帮助？

所有实验均使用 Qwen2.5-Omni，采用表11中的 LoRA 设置、附录 D 中的奖励以及 1,000 个训练步。R0 是表3中的实验。**R1** 的每个批次中，一半取自干净训练对，另一半取自其仅声音版本：编解码器、音色库替换、移调，以及针对音色的旋律和调的新组合（表 5）。每个版本都保留其条目的问题和标签。**R2** 采用监督微调。它将 GRPO 损失替换为针对正确字母的交叉熵，同时保留提示词、批次、训练步数、LoRA 和优化器，因此看到的训练示例数量相同。房间回声、均衡器和嘶声从不用于训练。

表 27 展示了结果。监督微调呈现出与 GRPO 相同的模式。漏判率从 36.7% 降至 3.0%（下降 33.7 个百分点），但错误翻转仅从 41.0% 降至 27.8%（下降 13.2 个百分点）。对于 GRPO，降幅分别为 34.8 个百分点和 14.9 个百分点（表16）。每次声音变化对应的错误翻转次数也与 R0 几乎相同。因此，这种模式并不取决于训练算法：只要模型仅使用干净配对进行训练，就会出现这种情况。

加入仅声音配对（R1）后，房间回声条件下的错误翻转率从 16.6% 降至 4.5%，嘶声条件下则从 52.4% 降至 37.7%，尽管 R1 从未见过这两种情况。在移调条件下，错误翻转率并未下降，尽管 R1 使用了移调数据进行训练（两次运行均为 45.2%）。R1 仍将 30 对移调后的“相同”旋律对中的 25 对判为不同（R0：同样为 25 对），并将 32 对移调后的 “ 相同 ” 和弦对中的 30 对判为不同（ R0：31 对）。其翻转 / 保持准确率从 86.0% 上升至 88.2%。代价是干净准确率下降 3.4 个百分点，主要体现在和声上（从 96.8% 降至 84.4%）。在 Slakh2100测试集（表 23）上，R1 的翻转/保持准确率为 77.1%，而 R0 和 R2 分别为 72.3% 和 70.9%；在真实测试集（表 40）上，R1 为 70.1%，而 R0 和 R2 分别为 62.8% 和 62.9%。在这两个测试集上，大部分提升都来自音色（Slakh2100 上的干净音色准确率从 64.4% 升至 96.4%，真实录音上则从 63.3% 升至 84.7%）；而在真实录音中，嘶声条件下的错误翻转率并未下降（从 20.9% 升至 22.8%）。因此，在未见过的回声和嘶声条件下，渲染片段上的性能提升并不能说明模型会忽略真实音频中的声音变化。

表 27：Qwen2.5-Omni 的训练基线（LoRA，表 11 设置，1,000 步，渲染测试，%）。旋律到全部：干净准确率；均衡器到训练集：每次保持变化的错误翻转率（表 19 格式）。房间回声、均衡器和嘶声从不用于训练。R1 加入训练集划分的仅声音版本（编解码器、音色库替换、移调；对于音色，还包括新的旋律和调；表 5），占每个批次的一半，且每个版本都使用其项目标签。R2 将 GRPO 损失替换为基于标准答案字母的交叉熵；提示词、批次、步数、LoRA 和优化器均相同。

Run	Mel.	和声	Tim.	Rhy.	全部	准确率FS	MR	FFR	均衡器	房间	嘶声	Tr.
不训练	68.8	59.2	66.8	50.0	61.2	61.2	36.7	41.0	33.7	37.2	49.0	48.4
R0：GRPO，干净（表3）	99.6	96.8	99.6	100.0	99.0	86.0	1.9	26.1	1.4	16.6	52.4	45.2
R1：GRPO，干净 +仅声音	100.0	84.4	98.0	100.0	95.6	88.2	4.9	18.7	2.7	4.5	37.7	45.2
R2：监督微调，干净	100.0	89.6	100.0	100.0	97.4	84.6	3.0	27.8	2.4	20.7	52.7	46.0

L 冻结的音频编码器与嘶声

经过 GRPO 训练后，Qwen2.5-Omni 在干净配对上的得分为 99.0%，但仍会对轻微的嘶声作出反应。本附录将这种行为与冻结的音频编码器联系起来（第5.3节）。

干净音频对。 与标准的相同/不同设计一样，每个干净的旋律、和声和音色“相同”配对中的片段 A 和片段 B 都完全相同（共 401 对：分别为 135、124 和 142 对），而在全部 349 对“不同”配对中，两者均有所不同。因此，干净准确率无法区分比较音乐本身的模型与对任何声音差异都会作出反应的模型。翻转/保持项目体现了这一区别，下面的分析将探究训练后的模型属于哪一种。

基于冻结编码器的规则。 在每次运行中，音频编码器都被冻结（表11）。因此，训练只能改变语言模型读取编码器输出的方式。为了了解模型学会读取什么，我们仅基于编码器构建了一条规则。我们分别使用

Qwen2.5-Omni编码器对片段 A 和片段 B 进行编码，并逐帧计算两个输出之间的平均余弦距离。当该距离高于阈值时，这条规则回答“不同”；每项任务分别使用一个阈值，并在干净训练集的训练对上拟合。该规则在这些配对上的得分为100%，在干净测试集的测试配对上为99.7%（表28）。在保持不变的配对上，训练后的模型与该规则给出相同答案的题目占85.1%，而训练前为56.0%。在嘶声配对上的一致率为98.6%，在移调配对上为98.4%，但在均衡器和房间回声配对上分别只有81.4%和67.1%。对于这两种变化，训练后的模型明显优于该规则（在均衡器条件下，错误翻转率分别为1.9%和18.6%；在房间回声条件下分别为22.7%和55.6%；表28），因此，训练教会模型的并不只是对编码器距离应用阈值。在嘶声和移调条件下——这两者会使编码器输出发生较大变化（如下）——模型的回答与该规则一致。

每种变化对编码器的影响有多大。表29比较了每种保持变化造成的距离与真实音乐变化造成的距离。嘶声造成的输出变化是实际变化的2.75倍，是均衡器造成变化的76倍。仅看旋律时，嘶声造成的变化是移动音符的12倍。对编码器而言，轻微的嘶声造成的变化比音乐本身还大。这个距离还能预测训练后的模型会把哪些“相同”的保持不变配对判断错：曲线下面积为0.92（401对中有218次错误翻转），而训练前为0.74。

编码器确实保留了一些有助于忽略嘶声的信息。在线性探针使用冻结特征（表28）并以干净配对进行训练时，嘶声条件下的错误翻转率为55.0%。加入房间回声、均衡器和嘶声版本的训练项目后，这一比例降至32.5%（均衡器：两种情况下均为0.0%）。简单的读出层可以学会忽略部分嘶声，但前提是嘶声出现在训练数据中。

嘶声扫描。嘶声要多轻，模型才会忽略它？我们以与渲染测试中嘶声保持变化相同的方式，在全部 1,000 个干净测试项中加入了 -80 至-30dBFS 的嘶声（表30）。我们只在片段 B 中加入嘶声，或在两个音频片段中加入相同的嘶声。仅在片段 B 中加入嘶声时，训练后的模型在 -80dBFS（我们测试的最低电平）下，就已将 42.9% 的“相同”配对判为不同（不加嘶声时为 2.0%）。从 -50dBFS 起，它会将其中 98% 至 100% 判为不同。在两个音频片段中加入相同的嘶声时，“相同”配对中的两个音频片段仍完全一致，但模型依然会将其中 7.5% 至 51.6% 判为不同。因此，大多数错误翻转源于两个音频片段之间的声音差异，但嘶声本身也会让模型倾向于判为“不同”。在节奏测试中，片段 B 加嘶声时的准确率从 -80dBFS 下的 98.8% 降至 -50dBFS 下的 53.2%。从未接触过嘶声的 R1 则能维持更久：在 -80 dBFS 下，它会将 0.7% 的“相同”配对判为不同；在 -60dBFS 下，这一比例为 43.9%。

我们还在 115 对“相同”音频片段的子集上运行了 Gemini 2.5 Pro（表31）。当片段 B 添加 -70 或-50dBFS 的嘶声时，其出错频率与不添加嘶声时大致相同。在 -30 dBFS 时，它在 40.0% 的情况下出错，而训练后的 Qwen2.5-Omni 的错误率为 100%。因此，Gemini 2.5 Pro 比训练后的模型能容忍响亮得多的嘶声，但它也开始发生翻转。

表28：基于冻结的 Qwen2.5-Omni 音频编码器构建的规则，与模型本身在渲染测试集的旋律、和声和音色任务上的表现进行对比（750 个干净项目、1,500 个翻转/保持项目；%）。距离：片段 A 和片段 B 分别独立编码后，编码器输出之间逐帧余弦距离的均值。阈值：每个任务各设一个阈值，根据干净训练对拟合（高于该阈值判定为“不同”）。探针：对帧特征绝对差值均值进行逻辑回归；第二个探针加入训练项目的房间回声、均衡器和嘶声版本（7,500 个干净项目和 9,000 个此类音频片段对）。均衡器至训练集：每次保持变化的错误翻转率。

	干净 $\mathrm{Acc}_{\mathrm{FS}}$		MR FFR		均衡器	房间	嘶声	Tr.
编码器距离上的阈值（未训练）	99.7	77.9	0.5	43.6	18.6	55.6	55.0	46.8
线性探针，干净训练对	99.7	80.7	0.5	38.0	0.0	54.1	55.0	46.8
线性探针，+回声、均衡器、嘶声版本	100.0	88.5	0.0	23.1	0.0	23.7	32.5	45.2
Qwen2.5-Omni，未训练	64.9	63.3	32.9	40.4	29.5	37.2	49.8	48.4
Qwen2.5-Omni、GRPO LoRA	98.7	84.1	2.5	29.2	1.9	22.7	53.6	45.2

表29：相对于真实变化，各种保持变化会使冻结编码器的输出偏移多远：在具有该变化的“相同”保持不变配对上测得的中位编码器距离，除以不同的变化配对（同一任务集）上的中位距离。数值大于 1 表示声音变化引起的输出偏移大于音乐变化。音色没有移调保持不变的配对。

保持变化	全部三个任务旋律		和声	音色
EQ	0.04	0.14	0.03	0.02
房间回声	0.93	3.39	0.97	0.50
Hiss	2.75	11.98	2.76	1.28
移调	0.65	2.11	1.18	–

表30：对全部 1,000 个干净测试项进行嘶声扫描（%），处理方式与渲染测试中的嘶声保持变化相同。左图和中图：在 401 对旋律、和声和音色“相同”的配对中，回答为“不同”的比例（0% 为正确），分别对应仅在片段 B 上添加嘶声，以及在两个音频片段上添加相同嘶声的情况。右图：在片段 B 上添加嘶声时的节奏准确率（250 个题目）。基底：未经训练的 Qwen2.5-Omni；R0：GRPO LoRA（表3）；R1：使用仅声音配对进行 GRPO 训练（附录 K），训练过程中从未接触嘶声。

	B片段加嘶声：“相同”配对			两个片段均加嘶声：“相同”配对			B片段加嘶声：节奏准确度		
嘶声（dBFS）	基底	R0	R1	基底	R0	R1	基底	R0	R1
none	17.7	2.0	0.0	17.7	2.0	0.0	50.0	100.0	100.0
−80	48.6	42.9	0.7	35.7	7.5	0.0	49.6	98.8	99.2
−70	60.1	51.1	10.2	42.4	23.7	4.7	51.6	92.8	98.8
−60	66.3	86.5	43.9	55.4	50.6	12.2	56.0	66.4	89.2
−50	70.6	98.3	73.6	53.9	29.4	7.5	55.6	53.2	68.0
−40	72.1	100.0	77.3	60.1	46.1	10.5	54.0	51.6	63.2
−30	72.8	100.0	83.3	64.8	51.6	30.7	52.8	53.6	64.0

表31：Gemini 2.5 Pro 在嘶声扫描的一个子集上的表现：每个级别均使用相同的 115 组旋律、和声和音色的“相同”配对（回答“不同”或未给出答案字母的比例；0% 为正确）。无答案字母：其中未给出答案字母的回复，计为错误。 p：与不含嘶声的相同配对（表2中的测试）进行精确麦克尼马尔检验。两个 Qwen2.5-Omni 列使用的完全是相同配对。

嘶声（dBFS）	Gemini 2.5 Pro 不加字母		p对比无字母	Omni、GRPO Omni、基底	
none	2.6		–	2.6	18.3
−70, on B	4.3	1	0.69	52.2	66.1
−50, on B	7.8	4	0.15	100.0	76.5
−30, on B	40.0	2	$< 10^{-10}$	100.0	76.5
−70, 两者都采用	4.3	4	–	20.0	43.5
−50, 两者都采用	0.9	0	–	28.7	56.5
−30, 两者都采用	9.6	3	–	51.3	69.6

M 配对指标与逐任务显著性

第 3.3 节指出，FFR 也会统计模型在干净音频对上已经犯下的错误。表 32 重新计算了表 2 和表 3 中每个模型及每次运行的配对指标。

这一差异对 Gemini 2.5 Pro 很重要。在旋律、和声和音色的 750 个保持项中，它答错了 209 题，其中 155 题是对实际存在差异的配对回答“相同”（表21）。这些都是错失的音乐变化，与它在存在差异的干净配对中有 41.0% 答错属于同一类错误，并非对声音的反应。若只统计干净版本答对的题目，其错误翻转率便从 26.9% 降至 20.8%。对训练后的 Qwen2.5-Omni 来说，同样的处理几乎不会改变错误翻转率（从 26.1% 降至 25.7%）。它的错误翻转是在没有声音变化时本来答对、却新出现的错误。它会在 97.1% 的翻转版本上改变答案，这正是预期的；但在 25.7% 的保持版本上也会改变答案。按正确答案划分，在答案为“相同”的保持不变的配对中，它有 54.4% 答错；在答案为“不同”的配对中，答错率为 0.3%。Gemini 2.5 Pro 的对应比例为 13.5% 和 44.4%。将 Gemini 2.5 Pro 未用字母作答的回复计为错误，最多会使其所在行的数值变动 1.1 个百分点。

表 33 测试了 GRPO 前后的各项任务。在 32 个运行与任务的组合中，有 22 个结果达到显著水平，且全部都是提升。正如第 5.2 节所述，节奏表现仅在 Qwen2.5-Omni（两次运行均如此）以及采用全量微调的 Audio Flamingo 3 上有所改善。我们不对节奏方面的其他变化作出任何判断。

表 32：渲染测试上的配对指标，根据表 2 和表 3 的运行结果重新计算（%）。FFR_{same} / FFR_{diff}：在答案为“相同”（401）/“不同”（349）的保持不变的配对中，旋律、和声和音色的错误翻转率。FFR_c =P（保持错误 | 干净正确），MR_c =P（翻转错误 | 干净正确）。 Δ保持 / Δ翻转：答案从干净版本变为保持版本 / 翻转版本的题目所占比例（理想值为 0% / 100%）。条目：干净、翻转和保持版本均回答正确（猜测：12.5%）。商业模型：“如 A”会将没有答案字母的回复按表 2 记为“A”；“无字母 =错误”则将其计为错误。这类回复数量：GPT-Audio-mini 为 1,385，GPT-Audio-1.5 为 241，Gemini 2.5 Pro 为 21，Gemini 3.8 Flash 为 0（共 3,000 条）。开放模型始终会给出一个字母。 †第500步。

模型	GRPO/ 评分	FFR	FFR_{same}	FFR_{diff}	FFR_c	漏检率	漏检率。	Δ保持	Δ翻转	条目
Qwen2-Audio	none	50.9	97.3	0.3	15.0	47.7	85.4	16.1	14.3	3.3
	LoRA	42.7	77.6	1.1	30.5	29.7	34.3	30.4	51.4	28.1
	full	45.4	82.0	0.9	29.3	31.7	39.1	27.2	46.1	24.4
Qwen2.5-Omni	none	41.0	40.4	40.4	40.4	36.7	42.6	47.2	45.7	21.0
	LoRA	26.1	54.4	0.3	25.7	1.9	1.9	25.7	97.1	72.2
	full	28.9	60.8	0.3	26.1	4.2	4.4	25.1	92.0	67.1
Audio Flamingo 3	none	47.6	5.7	93.7	17.8	49.6	79.9	18.5	18.1	5.8
	LoRA	41.4	60.3	11.2	31.7	29.7	33.0	32.9	53.6	31.8
	全量†	36.7	64.3	3.2	28.3	22.5	23.6	28.9	64.5	41.9
Phi-4-multimodal	none	48.3	5.5	94.3	19.7	53.5	83.6	20.0	18.4	3.7
	LoRA	48.4	1.0	99.7	14.3	53.4	88.1	14.3	13.5	3.0
	full	39.3	24.7	46.4	30.8	42.9	48.3	36.8	45.0	22.6
Kimi-Audio	none	45.7	8.2	88.3	22.6	52.3	76.5	25.4	24.4	7.5
MiMo-Audio	none	46.1	5.2	92.6	18.1	49.4	79.6	17.8	15.9	5.8
Fun-Audio-Chat	none	46.1	34.9	56.7	34.3	46.6	60.4	37.3	35.6	13.0
Step-Audio 2	none	45.5	34.9	55.6	33.3	49.1	61.0	36.3	37.1	11.9
GPT-Audio-mini	作为“A”	50.3	28.7	73.4	41.1	50.6	64.0	40.6	36.3	9.1
	没有字母 =错误	73.5	60.3	73.4	63.4	72.2	67.7	33.4	63.1	2.8
GPT-Audio-1.5	作为“A”	50.1	89.3	4.9	20.5	45.8	74.0	22.0	23.1	6.5
	没有字母 =错误	52.3	92.8	4.9	16.9	51.3	85.7	19.4	20.8	3.2
Gemini 2.5 Pro	作为“A”	26.9	13.5	44.4	20.8	24.7	28.5	26.9	58.7	43.6
	没有字母 =错误	27.4	14.2	44.4	21.4	25.2	29.1	27.6	58.2	42.5
Gemini 3.8 Flash	作为“A”	39.1	52.1	22.9	32.8	33.0	31.7	38.0	57.3	30.1
	无字母错误	39.1	52.1	22.9	32.8	33.0	31.7	38.0	57.3	30.1

表 33：GRPO 前后每项任务的干净准确率（%，每个单元格包含 250 个题目）。 *：在精确麦克尼马尔检验中显著，并对 32 个运行与任务对进行霍尔姆校正（$p_{\text{Holm}} < 0.05$）。32 个比较中有 22 个达到显著性，且全部为提升。 †第500步。

模型	GRPO	Mel.	Har.	Tim.	Rhy.
Qwen2-Audio	LoRA	46.0→70.8*	50.4→60.8*	43.2→92.4*	49.2→51.6
	full	46.0→70.4*	50.4→61.6*	43.2→80.8*	49.2→52.0
Qwen2.5-Omni	LoRA	68.8→99.6*	59.2→96.8*	66.8→99.6*	50.0→100.0*
	full	68.8→100.0*	59.2→85.2*	66.8→100.0*	50.0→99.6*
Audio Flamingo 3	LoRA	51.2→57.2	48.8→74.8*	58.0→97.6*	52.8→50.8
	全量†	51.2→68.0*	48.8→80.8*	58.0→99.6*	52.8→69.2*
Phi-4-multimodal	LoRA	54.4→52.0	47.2→49.6	56.8→56.8	50.8→50.8
	full	54.4→55.6	47.2→54.0*	56.8→88.8*	50.8→50.8

N 编辑后的片段播放两次，并根据变化幅度进行检测

大幅度的和声编辑可能会产生不寻常的和弦。将三度音移动 300 音分后，会得到根音、小二度和五度，或根音、三全音和五度。此时，模型可能会回答“不同”，因为编辑后的片段单独听起来很奇怪，而不是因为它比较了两个音频片段。为检验这一点，我们为每个旋律和和声测试项构建了两种“相同”配对：将编辑后的片段播放两次，以及将原始片段播放两次（表34）。每对中的两个音频片段完全相同，因此正确答案始终是“相同”。如果模型会对奇怪的声音作出反应，就会将编辑后的配对判为不同。

训练后的 Qwen2.5-Omni 没有显示出明确效果。对于编辑后的片段和原始片段，它回答“相同”的频率大致相当（和声在 300 音分时分别为 100.0% 和 96.4%；在全部和声级别上分别为 90.4% 和 94.8%；旋律在 700 音分时分别为 100.0% 和 100.0%）。因此，仅凭编辑后的片段并不会让它回答“不同”。对照实验无法揭示两个片段不同时模型会使用哪些线索。这里的两个片段完全相同，因此模型可以通过识别片段相同而通过测试，但在其他配对中仍可能利用编辑后片段中的线索。其他训练后的模型确实表现出效果。经过 GRPO 训练后，Qwen2-Audio 和 Audio Flamingo 3 在编辑后的片段播放两次时，在旋律和和声任务中回答“相同”的频率都降低了。Audio Flamingo 3（全量）对编辑后的 300 音分和弦回答“相同”的比例为 27.7%，而对原始和弦为 72.3%；对 700 音分旋律的相应比例为 17.7% 和 46.8%。Qwen2-Audio（LoRA）的比例分别为 0.0% 和 30.1%，以及 25.8% 和 50.0%。对于这两个模型，部分提升可能来自编辑后的片段本身。Gemini 2.5 Pro 在和声任务中也表现出同类但较小的效果，在旋律任务中则没有。在 300 音分时，它对编辑后的和弦回答“相同”的比例为 83.1%，而对原始和弦为 96.4%（83 个题目中分别有 69 个和 80 个，精确 McNemar检验 $p = 0.007$）。在全部级别上，这两个比例分别为 88.4% 和 98.4%（$p < 10^{-5}$）。对于 700 音分旋律，两种情况下它回答“相同”的比例均为 96.8%。

变化达到 100 音分时，听起来都不奇怪：大三和弦会变成小三和弦，反之亦然，而这两种都是常见的和弦。表35列出了不同变化幅度下的检测率。训练后的 Qwen2.5-Omni 能检测出所有 100 音分的和声变化（84 次中检测出 84 次），并能检测出各个变化幅度下 98.4% 至 100.0% 的旋律变化。训练前，它的检测率为 25.8% 至 46.0%。由于 100 音分的变化听起来并不奇怪，因此该幅度下检测到的 84 次和声变化不可能是因为和弦听起来奇怪。对于幅度更大的编辑，编辑后的片段中的线索也可能有所帮助。

表34：编辑后的片段播放两次（编辑）和原始片段播放两次（原始）：回答“相同”的比例（%；两个音频片段始终完全相同，因此100%为正确答案）。每个旋律和和声测试条目各包含一对相应类型的音频片段（每项任务250个：每个旋律等级62或63个，每个和声等级83或84个）。对于 Gemini 2.5 Pro，不含答案字母的回复计为非“相同”。 †第500步。

		旋律，全部		旋律，700 c		和声，全部		和声，300c	
模型	GRPO	Ed.	原始	Ed.	原始	Ed.	原始	Ed.	原始
Qwen2-Audio	none	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LoRA	29.6	54.4	25.8	50.0	9.2	20.8	0.0	30.1
	full	26.4	51.2	19.4	48.4	10.0	21.2	0.0	31.3
Qwen2.5-Omni	none	95.6	97.2	91.9	96.8	77.2	81.6	67.5	80.7
	LoRA	**100.0**	**100.0**	**100.0**	100.0	90.4	**94.8**	**100.0**	96.4
	full	**99.6**	**100.0**	98.4	100.0	64.8	68.0	86.7	69.9
Audio Flamingo3	无	87.2	88.0	85.5	85.5	99.2	**96.8**	**100.0**	97.6
	LoRA	12.0	24.4	6.5	19.4	86.8	97.2	67.5	96.4
	全量†	30.4	48.4	17.7	46.8	64.8	68.4	27.7	72.3
Phi-4-multimodal	none	98.8	**98.4**	**100.0**	96.8	85.2	87.6	86.7	86.7
	LoRA	97.2	97.6	96.8	**95.2**	**100.0**	**100.0**	**100.0**	100.0
	full	65.2	71.2	66.1	75.8	82.0	85.2	77.1	81.9
Kimi-Audio	none	89.2	89.2	91.9	88.7	95.2	93.2	96.4	94.0
MiMo-Audio	none	**100.0**	**100.0**	**100.0**	100.0	92.0	96.0	90.4	92.8
Fun-Audio-Chat	none	55.6	58.0	48.4	58.1	59.2	61.2	47.0	59.0
Step-Audio 2	none	84.0	84.0	79.0	85.5	40.8	45.6	34.9	49.4
Gemini 2.5 Pro	none	97.2	93.2	96.8	96.8	88.4	98.4	83.1	96.4

表35：按变化幅度检测真实变化（在干净和变化配对中，回答“不同”的比例；每个级别分别包含63、63、62、62组旋律配对和84、83、83组和声配对）。无论采用哪种评分规则，没有答案字母的回复都不算检测到变化。100音分的和声变化会将大三和弦变为小三和弦，或反之，因此两个和弦都很常见。图7汇总了干净、变化和保持版本，因此其中 Gemini 2.5 Pro 的数值与这里不同；这里的音频没有声音变化。 †第500步。

		旋律，音分				和声，音分		
模型	GRPO	100	200	400	700	100	200	300
Qwen2-Audio	none	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	LoRA	77.8	84.1	96.8	**98.4**	**100.0**	**100.0**	**100.0**
	full	84.1	85.7	96.8	**95.2**	**100.0**	100.0	100.0
Qwen2.5-Omni	none	46.0	39.7	25.8	29.0	38.1	47.0	42.2
	LoRA	**98.4**	**100.0**	**98.4**	**100.0**	**100.0**	**100.0**	**100.0**
	full	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Audio Flamingo3	无	15.9	15.9	9.7	9.7	3.6	0.0	2.4
	LoRA	90.5	90.5	**96.8**	**100.0**	34.5	42.2	86.7
	全量†	73.0	81.0	95.2	98.4	94.0	78.3	100.0
Phi-4-multimodal	none	0.0	0.0	1.6	0.0	9.5	12.0	10.8
	LoRA	0.0	1.6	0.0	3.2	0.0	0.0	0.0
	full	34.9	36.5	32.3	30.6	19.0	10.8	28.9
Kimi-Audio	none	11.1	17.5	14.5	16.1	6.0	8.4	8.4
MiMo-Audio	none	0.0	0.0	0.0	0.0	11.9	13.3	10.8
Fun-Audio-Chat	none	49.2	47.6	38.7	41.9	48.8	48.2	44.6
Step-Audio 2	none	15.9	12.7	14.5	11.3	53.6	61.4	56.6
GPT-Audio-mini	none	39.7	55.6	46.8	38.7	16.7	21.7	20.5
GPT-Audio-1.5	none	100.0	**95.2**	**100.0**	**100.0**	78.6	62.7	89.2
Gemini 2.5 Pro	none	66.7	58.7	50.0	51.6	39.3	41.0	55.4
Gemini 3.8 Flash	none	71.4	84.1	80.6	79.0	64.3	61.4	61.4

O 更长的旋律

我们的旋律只有四个音符。更长的旋律会改变结果吗？我们构建了包含 12 个和 24 个音符的旋律（每个音符 0.4 秒），每种设置有一个或两个音符发生变化，每种设置包含 200 个基础项目。这些旋律使用测试音高范围和音色库，并采用与旋律任务相同的翻转/保持构造。表36测试了现有检查点，未进行重新训练。

结果保持不变。模型的排名与 4 音符时的排名基本接近（在 16 个开放配置中，斯皮尔曼 ρ 0相关系数为 0.61 至 0.79，$p < 0.02$ 在每种设置中均如此）。除 Qwen2.5-Omni 外，未训练的开放模型仍接近随机水平（在翻转/保持任务上的准确率为 48.5% 至 56.8%）。对于训练后的 Qwen2.5-Omni 来说，更长旋律中的一个变化音符更难识别（LoRA：12 个音符时为 69.0%，24 个音符时为 61.8%，而 4 个音符时为 83.6%）。下降主要来自漏判（分别为 22.0% 和 38.0%，而 4 个音符时仅为 0.4%）。这与附录L相符：24 个音符中一个音符的变化对音频片段的影响较小，而训练后的模型主要是根据音频片段变化的幅度做出反应。错误翻转率仍然很高（40.0% 和 38.5%）。对于大多数能够听辨的模型来说，两个变化音符更容易识别。我们没有运行 Gemini 2.5 Pro：在 12 音符条目的第一批 75 个答案中，有 25 条回复没有答案字母，因此其得分衡量的将是缺失答案，而不是听辨能力。

表36：更长的旋律（%）。4 个音符：表13和表15中的旋律翻转/保持准确率。其他列：新题目，每个设置包含 12 或 24 个音符（每个 0.4 秒）和一个或两个变化的音符；每个设置有 200 个基础项目（200 个干净项、400 个翻转/保持项），构建方式与旋律任务相同（测试音高范围和音色库、相同的翻转/保持构造；保持项会改变均衡器、房间回声、嘶声和移调）。FS：翻转/保持准确率。使用现有检查点，未重新训练。n/r：未运行（在首批 75 个 12 音符题目的答案中，Gemini 2.5 Pro 有 25 条回复没有答案字母）。 †第500步。

			12 个音符，1 处变化		12 个音符，2 处变化		24 个音符，1 处变化		24 个音符，2 处变化	
模型	GRPO	4 个音符	干净	FS	Clean	FS	Clean	FS	Clean	FS
Qwen2-Audio	none	50.0	46.5	50.0	53.0	50.0	47.5	50.0	54.0	50.0
	LoRA	64.4	65.0	55.5	71.5	65.8	61.5	57.3	67.5	64.5
	full	61.8	56.0	51.3	64.0	60.5	55.5	56.3	60.5	60.5
Qwen2.5-Omni	none	60.0	58.5	49.8	60.5	63.8	54.0	53.5	64.5	66.0
	LoRA	83.6	83.0	69.0	95.0	83.5	67.5	61.8	84.0	79.0
	full	83.0	89.0	73.8	96.5	85.0	77.0	70.5	92.0	84.3
Audio Flamingo 3	无	51.6	52.5	49.0	49.0	50.3	56.5	49.5	48.5	49.3
	LoRA	54.4	56.0	53.8	59.0	57.3	56.0	52.3	62.5	54.0
	全量†	58.6	65.5	57.3	68.5	60.5	60.0	60.0	64.0	60.0
Phi-4-multimodal	none	49.6	53.5	50.0	47.0	50.0	52.5	50.0	46.0	50.0
	LoRA	48.6	53.5	50.3	46.5	50.0	52.0	49.8	48.5	49.8
	full	51.4	51.0	52.3	48.5	50.8	49.5	49.3	44.0	48.5
Kimi-Audio	none	52.8	55.0	48.5	48.5	49.5	51.0	52.5	50.0	51.8
MiMo-Audio	none	50.0	53.5	50.0	47.0	50.0	52.5	50.0	46.0	50.0
Fun-Audio-Chat	none	49.8	53.5	48.8	51.0	56.8	50.5	51.3	56.5	51.8
Step-Audio 2	none	47.8	51.0	49.5	47.0	53.0	48.5	51.3	47.0	50.5
Gemini 2.5 Pro	none	73.2	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r
斯皮尔曼 ρ（4 条注释，16 种开放配置）			0.61 ($p = 0.011$)		0.77 ($p < 0.001$)		0.78 ($p < 0.001$)		0.79 ($p < 0.001$)	

P 分级音色和相近乐器对

我们的音色任务使用来自四个不同家族的四种乐器，训练后的 Qwen2.5-Omni 在该任务上的得分为 99.6%。它是否也能听出音色中更细微的不同？三个新的子任务对此进行测试。每个子任务包含 250 个基础项（其中一半为“相同”，使用字节完全相同的音频片段；另一半为“不同”），以及渲染集的翻转/保持构造。两个音频片段在 2.5 秒的时间窗口内，都从测试音高范围中演奏相同的四音符动机。

亮度。 将带限锯齿波（加法合成）通过二阶 Butterworth 低通滤波器处理。一个音频片段的截止频率为 1 kHz，另一个音频片段的截止频率则高出 1/3、2/3、1 或 2 个八度，起音时间为 5 毫秒。保持变化仅包括房间回声和嘶声，因为均衡器变化会改变亮度。

起音。 截止频率为 2 kHz 的相同合成器。一个音频片段的起音时间为 5 毫秒，另一个则为 20、50、100 或 200 毫秒。起音时间和频谱质心是音色研究中发现的两个主要维度（McAdams 等，1995年）。亮度和起音只涵盖了合成器可改变内容的一小部分，因此

这两个子任务是迈向分级音色测试的第一步，并非全面的音色测试。

接近的配对。 小提琴与中提琴、中提琴与大提琴、小号与长号、原声钢琴与电钢琴，以及尼龙弦吉他与钢弦吉他，全部来自测试音色库。

结果。 未经训练时，Qwen2.5-Omni 在每个变化级别上都会将 19% 至 47% 的变化配对回答为“不同”，但也会将 23.2% 的相同配对回答为“不同”（表37）。它的答案并不随变化幅度而变化。经过 GRPO 后，它再也不会将相同配对称为不同。但它也几乎不会将任何渐变变化判断为不同的乐器：亮度变化在低于一个八度时为零，在两个八度时为 14.5%，而起音变化几乎为零。每个音色问题询问的都是乐器是否相同。因此，这些答案可能意味着，模型会将同一合成器更明亮或更柔和的版本判断为相同的乐器。接近的配对避免了这种歧义。在这些配对中，训练后的模型能够听出原声钢琴与电钢琴（62%）以及尼龙弦吉他与钢弦吉他（56%）的差异，但很少能听出小提琴与中提琴（10%）、小号与长号（4%）或中提琴与大提琴（0%）的差异（表38）。因此，它在音色任务上的 99.6% 反映的是区分乐器家族，而不是辨别更细微的音色差异。Gemini 2.5 Pro（每个子任务 100 个基础项）在 150 个相同配对中错答 1 个，且该回答没有字母标记。在亮度提高两个八度的配对中，它听出了 25 对中的 2 对，而在其他任何变化级别上都没有听出差异。

表37：分级合成音色（%）。亮度：带限锯齿波上二阶低通滤波器的截止频率，以 1 kHz 为基准，对比高出 1/3 至 2 个八度的频率。起音：5 毫秒对比 20 至 200 毫秒。步进列：被回答为“不同”的变化配对（干净和翻转）所占比例。相同项：在亮度、起音和接近配对中汇总的、被错误作答的相同干净音频对所占比例（0% 为正确）。每个子任务包含 250 个基础项目（每步 62 至 64 个变化配对、375 个相同项）；Gemini 2.5 Pro 在 100 个项目上测试（每步 25 或 26 个变化配对、150 个相同项），未提供答案字母的回复计为错误。†第500步。

		亮度，八度				起音，毫秒				
模型	GRPO	1/3	2/3	1	2	20	50	100	200	相同项
Qwen2-Audio	none	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.9
	LoRA	34.4	45.2	56.5	54.8	9.4	24.2	40.3	80.6	29.9
	full	35.9	41.9	51.6	61.3	28.1	46.8	71.0	98.4	38.9
Qwen2.5-Omni	none	29.7	22.6	25.8	41.9	34.4	46.8	19.4	40.3	23.2
	LoRA	0.0	0.0	0.0	14.5	0.0	0.0	0.0	0.0	0.0
	full	0.0	0.0	1.6	19.4	0.0	0.0	0.0	3.2	0.0
Audio Flamingo 3	none	1.6	0.0	0.0	1.6	1.6	0.0	4.8	11.3	1.9
	LoRA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	全量†	0.0	0.0	0.0	24.2	0.0	0.0	0.0	0.0	0.0
Phi-4-multimodal	none	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LoRA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	full	0.0	11.3	27.4	48.4	0.0	0.0	6.5	4.8	1.1
Kimi-Audio	none	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1
MiMo-Audio	none	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Fun-Audio-Chat	none	6.3	4.8	6.5	16.1	4.7	1.6	4.8	1.6	6.7
Step-Audio 2	none	28.1	33.9	22.6	25.8	26.6	24.2	22.6	37.1	24.5
Gemini 2.5 Pro	none	0.0	0.0	0.0	8.0	0.0	0.0	0.0	0.0	0.7

表38：测试音色库中的音色接近的乐器对（变化后的干净配对和变化配对中回答“不同”的比例，每对50题）以及每个新子任务的翻转/保持准确率（%）。保持变化：亮度任务中的房间回声和嘶声（均衡器变化会改变亮度）；其他任务中的房间回声、均衡器和嘶声。Gemini 2.5 Pro：每个子任务100个基础项目；“–”：该乐器对不在其子集中。 †第500步。

		音色接近的乐器对：检测结果					翻转/保持准确率		
模型	GRPO	小提琴/中提琴	中提琴/大提琴	小号/长号	钢琴/电钢琴	尼龙弦/钢弦	亮度	起音	音色接近
Qwen2-Audio	none	86.0	86.0	88.0	100.0	100.0	50.0	50.2	49.4
	LoRA	100.0	96.0	96.0	100.0	92.0	54.6	52.8	79.2
	full	100.0	94.0	100.0	100.0	82.0	53.6	54.2	71.6
Qwen2.5-Omni	none	42.0	24.0	20.0	50.0	30.0	53.2	58.4	55.0
	LoRA	10.0	0.0	4.0	62.0	56.0	54.4	51.4	60.0
	full	14.0	0.0	4.0	82.0	64.0	55.8	52.2	61.6
Audio Flamingo 3	无	0.0	0.0	0.0	0.0	0.0	49.4	52.0	50.2
	LoRA	56.0	4.0	88.0	84.0	76.0	50.2	50.4	78.4
	全量†	58.0	0.0	64.0	90.0	96.0	55.8	50.6	77.8
Phi-4-multimodal	none	0.0	0.0	0.0	0.0	0.0	50.0	50.0	50.0
	LoRA	0.0	0.0	0.0	0.0	0.0	50.4	50.0	49.8
	full	26.0	68.0	18.0	48.0	22.0	59.2	52.0	64.8
Kimi-Audio	none	2.0	0.0	0.0	2.0	2.0	50.4	50.0	49.8
MiMo-Audio	none	0.0	0.0	0.0	2.0	0.0	50.0	50.0	50.8
Fun-Audio-Chat	none	2.0	2.0	0.0	10.0	22.0	52.6	48.0	50.8
Step-Audio 2	none	14.0	22.0	62.0	22.0	30.0	47.8	50.8	51.6
Gemini 2.5 Pro	none	22.0	–	36.0	4.0	–	50.5	49.0	52.5

Q 真实录音

构建。 真实测试集是 MusicListenBench（第 3.4 节）的第二个测试集，用于评估在真实录音、真实房间、真实麦克风和真实演奏者条件下的听辨能力，且仅用于评估。其构建方式与渲染集相同：采用相同的四项任务、片段时长、1 秒间隔、响度（每段音频为 -23 响度单位）、问题和字母规则。保持变化也相同（均衡器、房间回声、嘶声；对于旋律和和声，还包括 0 至 200 音分的移调），并使用我们生成器的代码制作。每项任务包含 150 个基础项，每个基础项都有一个翻转版本和一个保持版本（总计 600 个干净项和 1,200 个翻转/保持项目）。

- **旋律**。 从 GuitarSet（Xi 等，2018年）中独奏片段的麦克风录音提取的 4 至 6 个音符乐句（2.5 秒）：原声吉他，六位演奏者。每根弦的音符标注确认同一时刻只演奏一个音符。在“不同”配对中，使用 Rubber Band 将第二个音符移调 100、200、400 或 700 音分。在“相同”配对中，同一个音符经过移调器，移调量为 0 音分。
- **和声**。 取自 GuitarSet 伴奏（comping）片段的 2 秒窗口。这些录音使用六通道拾音器录制，每根琴弦都有一条单独的干净音轨。每个窗口都从一次扫弦开始，其中标注的每个音符都属于一个大三和弦或小三和弦，且没有更早的音符延续到窗口中。我们只对演奏三度音的琴弦进行移调，分别移动 100、200 或 300 音分（向下移一个大三度或向上移一个小三度，与我们的生成器相同）。其他琴弦保持不变。因此，“不同”配对是一个真实的扫弦和弦与同一次扫弦、但三度音经过移动的版本进行比较；在“相同”配对中，演奏三度音的琴弦经过移调器时的音分值为 0。
- **音色**。 取自 Medley-solos-DB（Lostanlen 等人，2018年）的 2.5 秒片段，涵盖单簧管、失真电吉他、长笛、钢琴、次中音萨克斯管、小号和小提琴（不包含歌手类别）。每一对中的两个音频片段均来自不同的录音，无论是“相同”配对还是“不同”配对。因此，房间或麦克风无法泄露答案。
- **节奏**。 来自 MUSDB18-HQ 鼓声分轨的 6 秒窗口（Rafii 等，2019年），每秒包含一个起始点。我们使用 Rubber Band 将其时间拉伸至原来的 1.10 至 1.50 倍，并将其剪切为 4.0 秒。较慢的音频片段也会以 1.0 倍通过 Rubber Band 处理，因此两个音频片段的处理方式相同。两个音频片段来自同一个窗口。

标签检查。 标签来自每个数据集的标注，我们还在音频上对其进行了检查。在每一对“不同”的旋律对中，编辑音符与其目标的偏差都在 20 音分以内（使用 pYIN 测量的中位偏差为 0 音分）。在“相同”的旋律对中，音符保持不变。对于和声，在 100、200 和 300 音分处，音级中三度音所占的色度能量比例分别降至原始值的中位数 43%、52% 和 22%。在“相同”的旋律对中，该比例保持在 100%。对于节奏，在时间拉伸系数为 1.10、1.20、1.35 和 1.50 时，较快音频片段与较慢音频片段之间的起始计数比率的中位数分别为 1.08、1.19、1.31 和 1.40。

评分。 与附录 J 中一样，没有答案字母的回复计为错误。在翻转/保持项目中，Gemini 2.5 Pro 有 3.7% 的题目给出了此类回复（其中五次是调用超时三次），其余题目的得分为 76.6%。GPT-Audio-1.5 有 6.1% 的题目给出了此类回复，GPT-Audio-mini 则有 45.5%，大多是拒绝处理音频。在它们作答的题目中，这两个模型的得分分别为 49.7% 和 51.5%。我们通过 Google 的批处理接口运行 Gemini 3.8 Flash，使用的请求与渲染测试集相同（附录 C）。

结果。 未经训练时（表39），开放模型在翻转/保持任务上的表现接近随机水平（48.3% 至 55.5%）。只有 Audio Flamingo 3（55.5%）和 Fun-Audio-Chat（54.6%）显著高于随机水平。Gemini 2.5 Pro 再次成为表现最佳的模型（73.8%，渲染测试集上为 73.7%，Slakh2100 上为 75.3%）。Qwen2.5-Omni 的表现降至随机水平（52.1%，渲染测试集上为 61.2%）。在全部 12 个未经训练的模型中，其排名与另外两项测试均一致。单看这八个开放模型，它们的分数彼此相差不超过 7.2 个百分点，且都接近随机水平，因此它们之间的排名信息量不大（与渲染测试集的斯皮尔曼相关系数 $\rho = 0.32$ ）。旋律是最难的任务：没有开放模型或训练后的运行结果在该任务上的干净准确率超过 55.3%，而 Gemini 2.5 Pro 为 72.7%。

在使用我们合成配对进行训练的各次运行中（表 40），只有 Qwen2.5-Omni 的能力得以迁移。使用 LoRA 时，它在翻转 / 保持任务上的准确率从 52.1% 升至 62.8%（在相同的 1,200

道题上进行精确麦克尼马尔检验，$p < 0.001$）；采用全量微调时则升至 61.7%。Qwen2-Audio、Audio Flamingo 3 和 Phi-4-multimodal 的表现与随机水平相差不超过 6 个百分点。在训练中加入仅声音配对（R1），Qwen2.5-Omni 的准确率升至 70.1%（相较于 R0，$p < 0.001$）。在干净配对上进行监督微调（R2）的效果与 GRPO 相同（62.9%）。在 Slakh2100 配对上训练后，Qwen2.5-Omni 使用 LoRA 达到 75.0%，与 Gemini 2.5 Pro 持平（$p = 0.50$）。在这里，两种错误率的降幅相近（在合成配对上使用 LoRA 进行 GRPO：漏判率从 46.5% 降至 37.0%，错误翻转率从 49.3% 降至 37.3%）；与另外两个测试集上的情况一样，在所有训练后的 Qwen2.5-Omni 运行中，错误翻转率始终不低于漏判率。下面将考察模型在两段录音之间存在自然差异时出现错误翻转的情况。

第二次录音。真实录音能够检验渲染音频无法检验的内容。在 GuitarSet 的伴奏片段中，演奏者会多次扫奏同一和弦。同一和弦的两次扫弦总会有些许差异，但听者应该将它们判为相同（表41）。在干净配对上训练的 Qwen2.5-Omni，每次运行都能检测出几乎所有移调三度音程（41 个中有 37 至 40 个）。但每次运行也会将 42 个同和弦配对中的 27 至 35 个判为不同。在干净配对上训练后，两段录音之间的任何差异都容易被视为变化。使用仅声音配对进行训练时（R1），模型只将 42 个同和弦配对中的 7 个判为不同（与 R0 相比的精确麦克尼马尔检验，$p < 0.001$），并检测出 41 个移调三度音程中的 26 个。大多数未经训练的模型都处于两个极端之一：Qwen2-Audio 对全部 83 个配对都回答“不同”，而 Phi-4-multimodal 对其中 82 个回答“相同”。Gemini 2.5 Pro 则与训练后的模型表现相反：它将 42 个同和弦配对中的 37 个判为相同，却只检测出 41 个移调三度音程中的 12 个。Gemini 3.8 Flash 将 42 个同和弦配对中的 27 个判为不同。

表39：无训练的真实录音测试（%；此测试集上的表2和表15）。干净：每项任务150道题（全部共600道）。翻转/保持：每项任务300道题（$\mathrm{Acc}_{\mathrm{FS}}$中共1,200道）。未给出答案字母的回复计为错误；在翻转/保持项目中，这类回复占比为：GPT-Audio-mini 45.5%、GPT-Audio-1.5 6.1%、Gemini 2.5 Pro 3.7%、Gemini 3.8 Flash 0.0%。12个模型的翻转/保持准确率斯皮尔曼 ρ 相关系数：在渲染测试上为0.80（$p = 0.002$），在Slakh2100测试集上为0.91（$p < 0.001$）。

	干净准确率					翻转/保持准确率						
模型	Mel.	和声	Tim.	Rhy.	全部	旋律	和声	Tim.	Rhy.	$\mathrm{Acc}_{\mathrm{FS}}$	MR	FFR
Qwen2-Audio	49.3	50.0	50.0	44.0	48.3	50.0	50.0	50.0	48.3	49.6	48.5	52.3
Qwen2.5-Omni	44.7	52.7	62.7	48.7	52.2	47.7	49.7	58.7	52.3	52.1	46.5	49.3
Audio Flamingo 3	55.3	56.0	64.0	52.7	57.0	54.0	53.0	65.3	49.7	55.5	44.5	44.5
Phi-4-multimodal	50.7	49.3	50.0	56.0	51.5	50.0	52.3	50.0	51.7	51.0	50.2	47.8
Kimi-Audio	50.0	48.0	50.0	52.0	50.0	48.3	52.3	50.7	50.7	50.5	50.3	48.7
MiMo-Audio	50.7	46.0	54.0	60.0	52.7	50.0	51.3	52.7	54.3	52.1	49.3	46.5
Fun-Audio-Chat	48.0	58.0	56.0	56.0	54.5	49.3	53.0	57.7	58.3	54.6	45.2	45.7
Step-Audio 2	49.3	53.3	63.3	58.0	56.0	47.7	47.0	54.7	43.7	48.3	53.8	49.7
GPT-Audio-mini	48.0	26.7	36.0	4.7	28.8	45.3	30.7	32.3	4.0	28.1	71.5	72.3
GPT-Audio-1.5	49.3	42.0	54.7	38.0	46.0	47.3	43.7	53.0	42.7	46.7	53.2	53.5
Gemini 2.5 Pro	72.7	64.0	76.0	88.7	75.3	67.3	60.3	75.7	91.7	73.8	25.3	27.2
Gemini 3.8 Flash	65.3	52.0	65.3	60.7	60.8	54.3	52.0	68.0	54.7	57.3	42.0	43.5
始终选择“A”					51.7					50.4	51.5	47.7

表 40：真实录音测试集上的全部开放模型配置（%；未使用真实录音进行训练）。旋律至全部：干净准确率。均衡器至移调：每次保持变化的错误翻转率（移调）。R1：使用干净的 +仅声音渲染对进行 GRPO；R2：使用干净渲染对进行监督微调。†第500步。

模型	训练	旋律	和声	Tim.	Rhy.	全部	准确率$_{FS}$	MR	FFR	均衡器	房间	Hiss	Tr.
无训练													
Qwen2-Audio	none	49.3	50.0	50.0	44.0	48.3		49.6	48.5	52.3	74.3	48.7	33.5 49.3
Qwen2.5-Omni	none	44.7	52.7	62.7	48.7	52.2		52.1	46.5	49.3	62.1	43.4	40.3 48.0
Audio Flamingo 3	无	55.3	56.0	64.0	52.7	57.0		55.5	44.5	44.5	34.5	50.4	52.4 41.3
Phi-4-multimodal	none	50.7	49.3	50.0	56.0	51.5		51.0	50.2	47.8	25.2	52.2	66.0 53.3
在渲染划分上进行 *GRPO*（表*3*）													
Qwen2-Audio	LoRA	49.3	52.7	50.0	45.3	49.3		49.6	48.5	52.3	75.2	48.7	32.5 49.3
	full	49.3	52.0	50.0	44.0	48.8		50.1	48.0	51.8	73.8	49.6	32.0 49.3
Qwen2.5-Omni	LoRA	50.7	76.7	63.3	67.3	64.5		62.8	37.0	37.3	46.6	43.4	20.9 48.0
	full	48.0	76.0	64.0	66.7	63.7		61.7	38.0	38.7	46.6	46.9	22.3 49.3
Audio Flamingo 3	LoRA	50.0	60.7	60.0	56.0	56.7		53.9	47.8	44.3	53.9	49.6	29.6 50.7
	全量†	48.7	57.3	67.3	46.0	54.8		55.7	43.3	45.3	51.0	46.0	35.0 57.3
Phi-4-multimodal	LoRA	52.7	50.7	50.7	56.0	52.5		50.9	50.7	47.5	26.2	50.4	65.5 52.0
	full	50.7	47.3	52.7	57.3	52.0		51.7	49.8	46.8	31.1	52.2	55.8 57.3
在 *Slakh2100*划分 上使用 *GRPO*（附录 *J*）													
Qwen2-Audio	LoRA	46.0	52.7	50.0	44.0	48.2		50.0	48.7	51.3	71.8	49.6	32.0 50.7
	full	44.0	49.3	51.3	44.0	47.2		49.6	49.0	51.8	71.8	48.7	34.0 50.7
Qwen2.5-Omni	LoRA	54.0	76.7	84.7	84.7	75.0		75.0	23.0	27.0	22.8	38.1	18.4 45.3
	full	51.3	72.0	82.7	70.0	69.0		68.8	31.0	31.5	27.2	42.5	23.8 48.0
Audio Flamingo 3	LoRA	45.3	58.7	68.0	46.7	54.7		56.8	41.7	44.7	46.6	50.4	35.9 54.7
	全量†	50.0	50.7	67.3	54.0	55.5		56.9	43.0	43.2	42.7	51.3	36.4 50.7
Phi-4-multimodal	LoRA	52.7	50.7	50.7	56.0	52.5		50.9	50.7	47.5	26.2	50.4	65.5 52.0
	full	42.7	46.7	48.7	56.0	48.5		51.9	48.3	47.8	27.7	46.0	65.5 57.3
附录 *K* 的基线（合成配对）													
Qwen2.5-Omni	R1	54.0	70.7	84.7	68.7	69.5		70.1	28.2	31.7	30.6	38.9	22.8 48.0
	R2	49.3	74.0	61.3	65.3	62.5		62.9	35.8	38.3	48.5	43.4	21.4 49.3

表41：第二次录音：同一份 GuitarSet 片段中同一和弦指法的两个真实扫弦录音，共83对，两次录音间隔至少2秒。“相同”（42）：录音1与录音2进行比较。“不同”（41）：录音1与录音2进行比较，其中演奏三度音的琴弦移高100、200或300音分。准确率：准确率（%）。相同：“不同”：同和弦配对被回答为“不同”（0表示正确）。变更：检测到：移位三度音被回答为“不同”。100至300：按幅度检测到的比例（%；分别为14、14和13对）。训练：在渲染数据或 Slakh2100划分上进行 GRPO（LoRA 或全量）训练；R1：使用仅声音配对进行 GRPO；R2：使用干净配对进行监督微调。未包含答案字母的回答计为错误。

模型	训练	准确率：相同：‘不同’	已更改：已检测到		100	200	300
Gemini 2.5 Pro	none	59.0	42 项中的第 5 项	41 个中的 12 个	35.7	35.7	15.4
Gemini 3.8 Flash	none	44.6	42 个中的 27 个	41 个中的 22 个	35.7	64.3	61.5
Qwen2.5-Omni	none	62.7	42 个中的 18 个	28，共 41	64.3	71.4	69.2
	LoRA, 渲染	63.9	27，共 42	38，共 41	100.0	78.6	100.0
	full, 渲染	59.0	42 项中的 32 项	41 项中的 39 项	100.0	85.7	100.0
	LoRA, Slakh2100	63.9	42 项中的 29 项	41 项中的 40 项	100.0	92.9	100.0
	full, Slakh2100	62.7	42 个中的 27 个	41 个中的 37 个	100.0	71.4	100.0
	R1, 渲染	73.5	7 / 42	26 / 41	64.3	57.1	69.2
	R2（监督微调），渲染	55.4	35 / 42	39 / 41	100.0	85.7	100.0
Qwen2-Audio	none	49.4	42 个中的 42 个	41 个中的 41 个	100.0	100.0	100.0
Audio Flamingo3	无	54.2	42 个中的 6 个	41 个中的 9 个	14.3	14.3	38.5
Phi-4-multimodal	none	49.4	42 项中的第 1 项	41 项中的第 0 项	0.0	0.0	0.0
Kimi-Audio	none	54.2	42 项中的第 6 项	41 项中的第 9 项	21.4	21.4	23.1
MiMo-Audio	none	49.4	42 个中的 6 个	41 个中的 5 个	21.4	7.1	7.7
Fun-Audio-Chat	none	63.9	42 个中的 18 个	41 个中的 29 个	71.4	64.3	76.9
Step-Audio 2	none	48.2	42 项中的 30 项	41 项中的 28 项	57.1	78.6	69.2