arXiv:2602.04085v1 [cs.SD] 3 Feb 2026

BASS𝄞：面向音乐结构与语义推理的音频语言模型基准测试

Min Jang[1,♠], **Orevaoghene Ahia**[1,♠], **Nazif Tamer**[1]**Sachin Kumar**[2], **YuliaTsvetkov**[1],**Noah A.Smith**[1,3]华盛顿大学，[2]俄亥俄州立大学，[3]艾伦人工智能研究所♠表示贡献相同

摘要

音乐理解是一项复杂的任务，通常需要对音频的结构与语义要素进行推理。我们推出 BASS，用于评估音频语言模型在音乐理解与推理方面的表现，涵盖四大类别：结构分段、歌词转录、音乐学分析和艺术家合作。BASS 包含 12 项任务中的 2658 个问题，涉及 1993 首独特歌曲，涵盖超过 138 小时、来自多种音乐流派的音乐与曲目，旨在评估模型在真实场景中的音乐学知识与推理能力。我们评估了 14 个开源及前沿多模态语言模型，发现即使是最先进的模型，在结构分段和艺术家合作等高阶推理任务上仍表现吃力，而在歌词转录方面表现最佳。我们的分析表明，当前模型能够有效利用语言先验，但在对音乐结构、人声和音乐学属性进行推理方面仍有局限。BASS 提供了一个评估框架，可广泛应用于音乐推荐与搜索，并有望为音频语言模型的发展提供指引。[1]

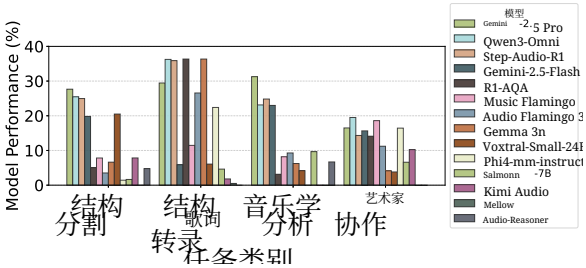


图 1：音频语言模型在 BASS 所有任务类别上的平均表现（每种情况下上限均为 100%）。表现最佳的模型在歌词转录任务上的得分最高，在艺术家合作任务上的得分最低。

1 引言

开发能够理解并推理多种音频领域内容的音频语言模型，近年来取得了显著进展，这些领域包括语音、环境声音和音乐（Goel etal.，2025年；Ghosh et al.，2025b；GeminiTeam et al.，2023年；KimiTeam et al.，2025年；Xu etal.，2025年；Xie etal.，2025年；Wijngaard et al.，2025年）。音乐尤其具有挑战性，因为它包含丰富的层次结构，涵盖和声、节奏、曲式等特征，而这些特征在较简单的音频领域中往往并不存在。准确的音乐理解能力能够让听众、学者、表演者、制作人、作曲家和媒体公司与单首音乐样本或大规模音乐合集中的内容进行开放式自然语言交互，从而使他们受益。

现有基准测试主要关注音乐描述生成或歌词描述（Manco 等人，2023年；Agostinelli 等人，2023年），或涵盖广泛的音频领域，其中包括音乐但并非专门针对音乐（Ahia等人，2025年；Sakshi 等人，2025年；Kumar等人，2025年；Ma等人，2025b；Huang 等人，2025年），因此无法评估模型在多样化情境下进行深入音乐推理的能力。MuChoMusic（Weck等人，2024年）弥补了其中一些不足，但完全依赖描述生成数据集（Manco 等人，2023年；Agostinelli 等人，2023年），使用的歌曲大多较短（~2分钟），且缺少要求进行长期结构理解或层次推理的任务。

我们推出 BASS，这是一项基准测试，专门评估音频语言模型在四大类别的 12 项任务中的音乐理解与推理能力：结构分割、歌词转录、音乐学分析和艺术家合作。我们的任务选择过程以音频语言模型在实际场景中的用例为指导，旨在回答以语言形式提出的真实世界音乐相关问题。结构分割包含需要识别

联系方式：minjang, oahia@cs.washington.edu **网站**：https://minjang10.github.io/bass-website/**代码**：https://github.com/minjang10/bass_音乐_基准测试。**数据**：https://huggingface.co/datasets/oreva/bass_音乐_基准测试

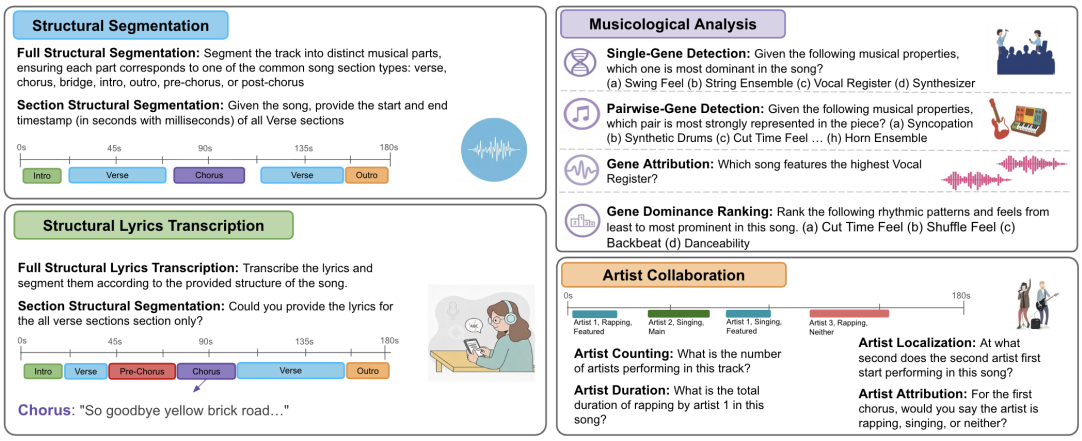


图 2：BASS基准中所包含任务的概览，旨在评估音频语言模型在需要长期结构或层次推理的音乐理解任务上的表现。该基准包含四大类别的十二项任务：结构分段、结构化歌词转录、音乐学分析和艺术家合作。

并定位录音中音乐段落的任务。歌词转录评估模型以不同粒度转录歌曲歌词的能力。音乐学分析侧重于识别抽象的音乐学属性并对其进行推理。艺术家合作包含需要识别多位演唱者、确定其身份并在歌曲中进行时间定位的任务，通常还涉及细粒度的时间推理。

我们使用 BASS 评估了 14 个开源及前沿多模态语言模型§3。结果表明，即便是 Gemini 2.5 Pro 这样的前沿模型，以及 Qwen3-Omni 这样的强大开源模型，在若干任务上的表现也明显吃力。如图 1所示，表现最佳的模型在歌词转录任务上成绩最好，在艺术家合作任务上表现最差。我们对所有任务中的模型性能进行了深入分析，结果表明，当前的音频语言模型在能够利用语言先验的任务上表现强得多，但在音乐结构、演唱风格和多艺术家档案方面的高阶推理能力仍然有限。BASS 为评估音频语言模型的音乐理解与推理能力提供了一个颇具挑战性的测试平台，凸显了它们当前的优势与局限。论文获接收后，我们将公开发布该数据集以及与此基准测试相关的全部评估代码。

2 BASS基准

BASS旨在评估人工智能系统的音乐理解与推理能力。它包含2,658组问答对，涵盖四大类别下的12项不同任务：a）结构分段 b）歌词转录、c）音乐学分析和 d）艺术家合作。与Weck等人 (2024年）等以往的音乐基准测试不同，后者主要使用包含四个干扰选项的多项选择题评估模型；BASS兼有开放式问题和多项选择题，部分题目的干扰选项增至八个，从而降低模型靠猜测完成任务的可能性，并能评估更深入的音乐推理能力。所有任务中的音乐录音均来源于YouTube。[2] 每个样本都经过人工核验，以确认歌曲准确无误，并确保音频保真度高。数据集中音频片段的平均时长为4.16分钟，范围从49.06秒到19.44分钟不等。附录中的表4提供了BASS的详细统计数据，下面我们将深入介绍每项任务严谨的标注流程。

2.1

结构分段是音乐信息检索领域中研究充分的一项任务，需要具备音乐知识，并理解音乐音频中的关键结构段落（Paulus等人，2010年；Ehmann等人，2011）。给定一段音乐录音，目标是识别重要的结构元素，并据此对录音进行时间分段。这项能力对于面向消费者的应用至关重要；Spotify 等流媒体平台会利用结构信息来改进推荐、歌单编排和精彩片段提取（

[2]由于许可限制，我们将提供这些曲目的 YouTube 链接，而不直接分发音频，从而在尊重版权的同时确保合法访问。

Montecchio等人，2020；Music Tomorrow，2025年；Choi等人，2016）。在这一类别中，我们纳入两项任务： **完整**和**分段**结构分段。这项任务的数据来源于 Harmonix Set（Nieto等人，2019），这是一个包含流行歌曲结构标注的数据集。该数据集最初包含 61 种结构段落标签，其中 5 种是非常常见的段落（前奏、主歌、副歌、桥段或尾奏），其余 56 种则出现频率较低。

完整结构分段 （FSS）要求模型将整首歌曲分割为其主要段落，并返回每个段落的起止时间戳。乐段结构分割 （SSS）针对每首歌曲的单一段落类型（例如副歌），要么返回该段落类型所有出现位置的时间戳（请注意，段落经常重复），要么返回录音中的特定段落（例如第二个副歌）。

2.2 结

结构化歌词转录在标准歌词转录（Wang 等人，2022；Mesaros 与 Virtanen，2008）的基础上加以扩展，要求模型转录歌曲每个结构段落中的歌词。我们使用 Genius [3] 歌词，因其采用协作式标注流程，热门曲目的转录文本和分段由数百名贡献者不断完善，从而确保更高的可靠性。该任务比单纯的结构分段更具挑战性，因为模型既必须对歌曲进行分段，又必须将生成的歌词正确对齐到每个段落。成功的表现可支持内容索引、音乐推荐，以及更可靠的自动化与本地化歌词转录等应用。在结构化歌词转录类别中，我们同样设有**完整**与**段落**变体。

完整结构化歌词转录（FSLT）：模型的任务是转录歌曲中每个段落的歌词，并以从 Genius 歌词抓取的歌词作为黄金标准参考。这些歌词包含常见段落（前奏、主歌、副歌、桥段、尾奏）。我们还纳入了经常出现的其他段落，例如预副歌和后副歌。为减少歧义，每个问题都采用模板格式：提示中列出所有相关段落，并要求模型生成每个段落对应的歌词。

段落结构化歌词转录（SSLT）：模型需要转录特定段落的歌词（例如桥段）。此任务包含两类问题：全段落级 SSLT 和单段落级 SSLT。对于全段落级 SSLT，模型会获得一个段落名称，并必须转录歌曲中该段落的所有实例。对于单段落级 SSLT，会提供一个特定实例（例如第二段主歌），模型只需转录该实例。

2.3

音乐学分析考察音乐的结构和表现元素——节奏、和声、旋律、音色、人声和歌词——为音乐分类和推荐等其他常见任务提供可解释的基础（Castelluccio，2006；Oramas 等人，2025年）。在此基准测试中，我们评估音频语言模型识别音乐学属性（称为基因）的能力：歌曲中具体且可衡量的属性，例如鼓的使用频率、电吉他的失真程度、悲伤歌词，或对节奏律动的强调。

我们从 MGPHot 数据集（Oramas 等人，2025年）中提取这些基因。该数据集包含 58 种音乐属性，分为七类：节奏、乐器编配、音响、和声、作曲重点、人声和歌词。这些属性由专业音乐学家作为音乐基因组计划的一部分进行标注（Castelluccio，2006），涵盖了 1958 至 2022 年间至少一次登上《公告牌》百强单曲榜的 21,320 首歌曲。音乐学家以 0–5 分制评定每个基因的主导性，之后将评分换算为 0.0–1.0 分制，其中 1.0 表示主导性高，0.0 表示不存在。基于这些专业标注，我们为音频语言模型精心设计了四项具有挑战性的任务：(a) **单基因检测**，(b) **成对基因检测**，(c) **基因归因**，以及 (d)**基因优势排序**。

单基因检测 （SGD）：模型需要根据每个基因的名称和描述，从四个选项中识别给定录音中最主要的音乐学属性。问题由该基因强烈表达的歌曲构建（≥ 0.6）。干扰项选自其取值与真实标注正相关、但彼此负相关的属性，以避免同一基因类别内出现过于相似的选项（例如，“合成器”和“合成器音色”）。

成对基因检测 （PGD）：要求音频语言模型

[3] https://genius.com/

从给定的录音中选出表现最为突出的那对音乐学属性，并提供八种音乐学属性及其描述。这项任务比 SGD 更具挑战性，因为真实属性对来自通常呈负相关的属性，例如（原声吉他和合成音色），但问题所依据的歌曲中，这两种属性都表现得十分突出。干扰项选自与至少一种真实属性呈正相关、但与所有其他干扰项均呈负相关的基因。

基因归因 （GA）将分析范围从单首歌曲扩展到评估四段录音中的音乐学属性。模型会得到一个音乐学属性及其描述，并须从四段录音中选出最能体现该属性的一段。符合条件的问题要求真实歌曲的音乐学属性值大于 0.8，而干扰歌曲的属性值必须低于 0.5，以确保有足够的对比度。模型会收到一个音频输入，其中包含依次拼接的每首歌曲，歌曲之间间隔 5 秒静音。模型应输出最突出地体现该属性的录音编号（第一段录音为 1，第二段为 2，依此类推）。评估通过计算四选一任务的准确率进行。

在 ，模型会收到每首歌曲的四个音乐学属性，并须根据它们的突出程度按升序排列。为避免歧义，每首歌曲的属性值都经过设计，彼此间隔均匀。模型必须严格按照正确顺序给出排序，才算正确。部分正确的输出不予计分。

2.4

多项研究分析了合作歌曲的音乐特征（Duinker，2021；Komaniecki，2017；Duguay，2022），但尚无研究分析音频语言模型对这些特性的理解效果。在流行音乐中，歌曲往往由多位演唱者共同呈现，例如主唱与特邀艺人，或由多名成员组成的乐队演绎。我们设计这一任务，以评估音频语言模型对艺人角色（主唱或特邀）、人声演绎类型（说唱或歌唱）及其在歌曲不同曲式段落（例如前奏、主歌、副歌）中时间分布的理解与推理能力。这包括模型定位每位艺人或声乐演绎类型在整首曲目中出现位置的能力。

问题与真实标注取自《协作歌曲数据集》（CoSoD；Duguay 等人，2023年），该数据集包含 331 首经过标注的多艺人歌曲，标注内容包括曲式结构以及叠录、声像定位和演绎类型等人声制作特征。我们设计这些任务，旨在评估艺人行为和人声表演的核心维度，具体包括艺人出现、时间安排、演唱风格以及特定段落中的艺人角色。这些任务涵盖了分析艺人在合作音乐中所做贡献时最相关的维度，可支持音乐推荐和自动生成元数据等实际应用。本类别包含四项任务，下面将逐一讨论。

艺术家计数 包含五个子任务。标准艺术家计数 要求模型报告一首歌曲中署名艺人的总数，不论其角色或贡献程度如何，主艺人和特邀艺人均包括在内。特邀艺人计数针对特邀艺人超过两位的歌曲，要求模型报告曲目中特邀艺人的总数。特邀艺人是指参与歌曲演出、但并非主唱艺人的其他表演者。声乐演绎计数要求模型统计单次录音中以特定声乐演绎风格（说唱、歌唱）演出的艺人数量。段落艺人计数是指统计歌曲某个特定段落（例如前奏、主歌、副歌）或多个段落中的艺人数量，也包括段落为空或不存在等否定情形。时序艺人计数要求统计给定时间范围内或多个时间戳处的艺人数量，也包括计数为零的越界区间。对于每项艺术家计数子任务，模型都应输出确切的艺人数量；评估时则将模型输出与真实计数进行比较，并采用精确匹配准确率作为指标。

艺术家时长包含四项子任务。目标艺人时长要求模型报告歌曲中特定艺术家的演唱总时长，以秒为单位。为确保对可能未曾听过这首歌的模型公平，艺术家按照出现顺序而非姓名来识别。声乐演绎时长要求模型报告歌曲中某种特定声乐演绎类型（例如说唱、歌唱）出现的总时长，以秒为单位，而不论由哪位

艺术家进行演绎。艺人演唱时长结合了前两项子任务，要求模型报告特定艺术家采用特定声乐演绎风格（例如说唱、歌唱）的总时长，以秒为单位。最后，段落时长要求模型估算特定段落类型（例如所有主歌乐段）的总时长，以秒为单位。评估采用精确匹配准确率，即模型输出的时长必须与真实时长相差不超过 ±3秒。

艺人定位 要求模型提供特定艺人在歌曲中首次出现的起始时间戳（以秒为单位）。特定艺人根据其在歌曲中的出现顺序来确定。我们使用真实标注时间戳前后各 A3 秒的时间窗口来评估每个模型的输出。评估采用精确匹配准确率：如果模型的输出落在真实标注时间戳的 ±3 秒窗口内，则视为正确。

艺人归属 包含四项子任务。 演唱风格比较 要求模型判断同一首歌曲中不同艺人的演唱风格相同还是不同。例如，任务可能会询问歌曲中的第一位和第二位艺人是否都在说唱。段落演唱风格 要求判断某位艺人在特定段落中是歌唱、说唱，还是两者皆非。 段落艺术家角色 要求模型判断某位艺人在某个段落中的角色（主唱、特邀艺人，或两者皆非）。例如，模型可能需要判断某位艺人在副歌中是主要表演者还是特邀艺人。最后， 时间演唱风格 要求模型识别某位艺人在给定起始和结束时间戳范围内的演唱风格。所有艺人归属任务均采用精确匹配准确率进行评估。

3 实验设置

模型 我们对广泛的音频语言模型进行基准测试。在开源模型中，我们纳入了 Qwen3-Omni（Xu 等人，2025年）、Audio Flamingo 3（Goel 等人，2025年）、Music Flamingo（Ghosh 等人，2025a）、Step-Audio-R1（Tian 等人，2025年）、R1-AQA（Li 等人，2025年）、Audio Reasoner（Xie 等人，2025年）、Phi 4 Multimodal Instruct（微软等人，2025年）、Voxtral（Liu 等人，2025年）、SALMONN- 7B（Tang 等人，2024年）、Mellow（Deshmukh 等人，2025年）、Kimi-Audio（KimiTeam 等人，2025年）以及 Gemma 3（Gemma Team 等人，2025年）。闭源-源模型包括 Gemini 2.5 Pro、Gemini 2.5 Flash（带思考）以及 Gemini 2.5 Flash（不带思考）（Gemini Team 等人，2023年）。[4]

评估策略 在我们的所有实验中，音频语言模型接收文本提示（指令）和音频文件作为输入，并生成文本作为输出。对于每项任务，作者手工编写了10个自然语言提示词，它们彼此互为释义，即都以略有不同的方式提出同一个根本问题。在每项任务的音频样本中，我们保持这些释义的均衡混合，以促进多样性。为确保模型真正处理音频并避免依赖捷径，除音乐学分析问题外，我们均采用自由生成方式评估所有任务；对于音乐学分析问题，我们使用多项选择格式，如§2.3所述。多项选择题上的表现通过四次运行的多数投票来确定。

评估指标 对于结构化歌词转录，我们使用词错误率（WER）。对于涉及多个段落的问题，参考段落与预测分段进行最优匹配，最终词错误率取所有段落的平均值。对于结构分段，使用交并比（IoU）衡量表现：仅当预测的分段标签与参考标签匹配时，重叠才计为交集。对于所有其他任务，我们使用完全匹配准确率（EMA）。对于需要时间戳或时长的问题，偏差在3秒以内的预测视为正确。

汇总各项任务的评估指标面临两项挑战。首先，不同评估指标的最优方向不同：词错误率（WER）越低越好，而 EMA 越高越好。此外，WER 没有上限。[5]其次，虽然 EMA 的取值范围是 0.0 到 1.0，但由于不同任务的随机猜测准确率各不相同，直接对不同任务的 EMA 取平均值并不合理。因此，我们对 WER 取反，使分数越高代表表演效果越好，即 $\mathrm{IWER} = \frac{1}{1+\mathrm{WER}}$，并将分数限制在 0.0 到 1.0 之间。为解决第二项问题，我们对 EMA 进行归一化，以表示模型相较于随机猜测的提升。具体而言，我们报告

[4]由于成本限制，我们不评估 OpenAI 模型；事实上，使用这些模型处理音频的成本大约是使用 Gemini 2.5 Pro 的 16 倍。[5]尽管标准自动语音识别任务的词错误率通常介于 0.0 和 1.0 之间，但由于音乐录音属于长音频，且包含更多背景噪声，我们发现许多模型的词错误率远高于 1.0，这尤其是由过多的插入造成的，因此词错误率没有上限。

		音乐学分析				结构分段		艺术家合作				歌词转录		平均值
模型	Size	SGD	PGD	GDR	GA	Full	段落	计数	时长	定位	归属	Full	段落	
Gemini-2.5 Pro	-	**64.98**	10.40	17.80	31.90	**32.34**	**22.98**	25.77	8.25	3.12	28.84	**42.63**	29.45	**26.54**
Step-Audio-R1	33B	43.88	8。89	14。62	**31.91**	32。19	17.73	19.34	6.42	**7.81**	23.74	21.86	35.89	22.24
Qwen3-Omni	30B	47.86	1.27	11.51	31.87	28.88	22.07	18.54	**9.78**	9.38	**40.32**	25.53	**36.25**	23.61
Gemini-2.5-Flash	-	43.48	4.06	**20.90**	23.50	24.47	15.15	20.40	7.66	3.12	31.44	5.13	5.89	17.10
Music Flamingo	8B	32.34	4.07	−4.34	0.67	13.07	2.59	33.54	4.27	4.69	31.88	15.72	11.47	12.50
R1-AQA	8.2B	12.04	−2.98	−4.35	7.87	5.91	4.16	**35.03**	0.92	0.00	20.51	40.05	36.35	12.96
Audio Flamingo 3	8B	25.60	8.01	−4.34	7.87	7.03	0.01	21.92	1.84	0.00	21.13	17.59	26.57	11.10
Gemma 3	5B	13.63	3.30	−1.19	9.22	8.92	4.35	12.25	0.30	1.56	2.47	40.05	36.35	10.93
Voxtral-Small	24B	16.82	5.42	5.15	−10.67	32.16	8.84	9.09	2.76	0.00	3.42	26.45	6.08	8.79
SALMONN-7B	7B	0.50	**37.26**	−4.03	4.93	3.18	0.12	27.32	0.61	0.00	−1.48	11.61	4.63	7.05
Phi4-mm-instruct	5.5B	−4.96	1.27	5.15	−4.93	2.71	0.16	23.59	3.66	0.00	38.51	4.99	22.45	7.71
Audio-Reasoner	8.4B	11.64	5.37	1.98	7.80	6.63	2.90	18.47	3.06	0.00	−59.03	2.46	0.00	0.00
Kimi Audio	7B	16.82	0.55	−4.35	−33.33	13.06	2.64	19.12	0.92	0.00	20.96	5.17	1.80	3.61
Mellow	167M	−32.94	−3.70	−4.35	−33.33	0.00	0.00	0.00	2.14	0.00	−62.50	53.50	0.55	−6.72

表1：所评估模型在 BASS 各项任务上的归一化表现，包括音乐学分析（单基因检测—SGD、成对基因检测—PGD、基因优势排序—GDR，以及基因归因—GA）、结构分段、艺术家合作和歌词转录。每项得分的上限均为 100%，粗体数值标示各任务中的最高分。归一化流程详见§3。原始得分见附录中的表6 。

$\text{EMA}_N = \text{EMA}_{1-R}^{-R}$ 其中 R 为随机猜测准确率。这样得到的上限为 1.0（但如果某个系统的平均表现低于随机猜测，EMA_N 可能为负值）。我们的所有结果均以百分比（%）报告。

4 结果与讨论

基准测试结果 在表1中，我们展示了所有模型在 BASS 全部任务上的归一化得分。原始得分见附录中的表6 。所有模型的平均表现仍低于 30%；最佳模型为 Gemini-2.5 Pro，得分为 26.54%，其次是 Qwen3-Omni（23.61%）和 Step-Audio-R1（22.24%）。值得注意的是，这些模型均采用混合思考模式进行训练，且评估时启用了思考功能。总体而言，模型在歌词转录方面表现最佳，而艺术家合作任务最具挑战性。性能差距表明，当前的音频语言模型在能够利用语言先验的任务上更为擅长，但在音乐结构、演唱风格和多艺术家档案方面的高阶定量及时序推理能力仍显不足。

下面，我们围绕指导本研究的研究问题，对结果进行详细讨论。具体而言，我们将考察音频语言模型如何理解和推理音乐结构与歌词、如何处理艺人角色和演唱风格，以及它们在不同音乐流派中的表现有何差异。

音频语言模型对音乐结构的理解和推理能力如何？ 在结构分段和歌词转录等结构任务中，我们观察到，与针对孤立段落的任务相比，当任务要求对完整歌曲进行时间定位时，模型往往表现更好。我们推测，按顺序分析较早的段落，能让模型利用前面片段中的上下文线索，更准确地识别后续段落。然而，即使模型对完整歌曲进行时间定位（见图 11），也经常会混淆声学上相似的段落，例如主歌和副歌。我们推测，这种混淆是因为音频语言模型并未经过专门训练，无法在较长上下文中理解音乐结构。这一点在我们对模型在不同歌曲段落上的性能分析中显而易见（见图 3）：模型通常更擅长定位歌曲中较早出现的内容，例如前奏和第一段主歌，而不擅长定位较晚出现的段落，例如桥段和尾奏。

提供自动语音识别转录文本是否能提升音频语言模型的结构化歌词转录？ 通用自动语音识别模型（Radfordet al., 2023年）在自动歌词转录方面展现出较强的泛化能力（Wanget al.,2023年; Zhuoet al., 2023年; Ou et al., 2022;Cífkaet al., 2024年; Radfordet al., 2023年）。然而，这些研究也指出转录文本中存在一些错误。这促使我们探索结构化歌词转录的级联方法。将转录文本中的语言线索与音频中的声学信息相结合，可能有助于模型更准确地确定段落边界并对齐歌词。我们首先使用 Whisper 转录歌曲，然后将所得转录文本与音频一并重新输入音频语言模型，作为执行结构化转录任务的引导。如表 2所示，

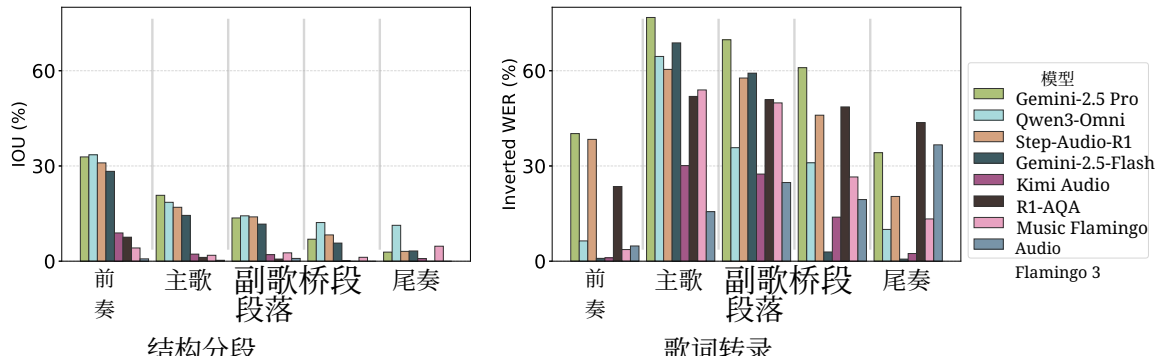


图 3：结构分段和歌词转录的分段表现（越高越好）。分割表现会随时间推移而下降，而转录表现（IWER）在前奏 中最低，相较于主歌和副歌段落。

我们发现，与纯音频转录相比，使用级联 Whisper 转录文本实际上降低了音频语言模型的歌词转录表现。我们推测，转录文本可能会误导模型并引入自动语音识别错误，从而损害需要精确时间定位的段落上的表现。

设置（IWER ↑）	Qwen-3 Omni	Gemini 2.5 Pro	Phi-MM Instr.	Gemma 3n
完整歌词转录（***FLT***）				
级联（Whisper+音频语言模型）	6.14	13.25	**3.02**	10.90
仅音频语言模型	**11.38**	**42.09**	0.90	**14.45**
分段歌词转录（***SLT***）				
级联式（Whisper+音频语言模型）	23.77	29.69	0.72	5.51
仅音频语言模型	**26.54**	**46.11**	0.3	**34.11**

表 2：与仅用音频进行转录相比，使用级联式 Whisper 转录文本会降低音频语言模型的歌词转录表现。（数值以反向词错误率报告；越高越好）

音频语言模型能否随时间变化，区分合作音乐中的艺人角色和演唱风格？ 合作音乐对音频语言模型而言是一项复杂挑战，模型在各任务类别中的表现都较低。在细粒度定位以及量化歌曲中各艺人贡献时长方面，表现尤其不佳。这一发现与先前研究相符；这些研究基于通用音频基准测试，考察时间定位（Ahia 等人，2025年；Kumar 等人，2025年；Ma 等人，2025b）。我们还观察到，模型识别说唱的能力通常优于识别歌唱（图 4）。这可能是因为说唱与口语更为相似，而模型通常接受过口语训练。

音频语言模型对音乐学属性的推理能力如何？ 音乐理解涉

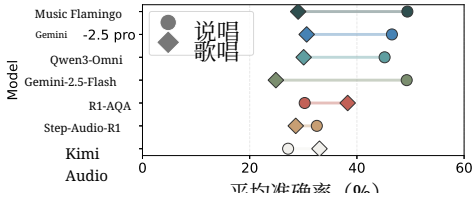


图 4：模型对歌唱与说唱演唱风格的识别准确率。大多数模型对说唱的识别比对歌唱更准确。

及对音频中存在的多种音乐学属性进行推理，从人声和歌词到节奏等。为评估音频语言模型对这些属性的捕捉能力，我们分析了不同音乐学属性上的模型表现，见图 5。结果因模型和属性而异。Gemini-2.5 Pro 总体表现最强，尤其在人声、歌词和作曲方面，而 Music Flamingo 主要在乐器编配上表现出色，但在和声方面落后。所有模型显然都难以应对音响和节奏，这表明捕捉细微的音色特征与时间细微差异仍是当前音频语言模型的关键挑战，这与我们此前观察到的音频语言模型在时间推理方面的困难相一致。

音频语言模型能否通过音源分离提升结构与音乐语义推理能力？ 音乐录音通常包含人声、乐器和背景声音的混合，与传统语音有显著差异。以往研究考察了音源分离——从混合录音中分离出各个音轨，如人声、贝斯、鼓和其他乐器——对下游音乐理解任务的影响（deBerardi-nis et al.，2020年；Xuet al.，2014年；Rosner and Kostek，2018年；Agrawal andKaramchandani，2023年；Syed

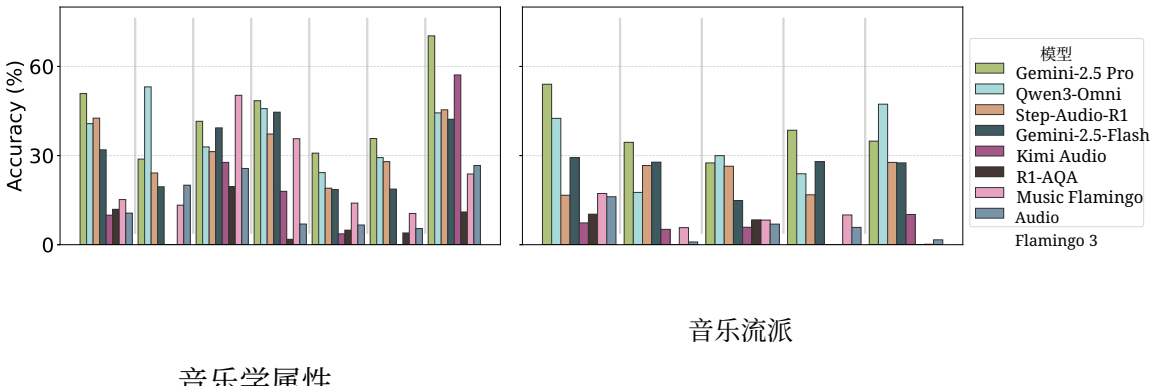


图 5：音乐学分析任务中，模型在不同音乐学属性和音乐流派上的表现。（**左**）各项具体属性的平均表现；模型在人声和歌词方面表现出色，但节奏、乐器编配和音响等属性，即使对前沿模型而言仍颇具挑战。（**右**）按音乐流派划分的表现情况，显示模型通常在乡村音乐和摇滚方面表现最佳，而在嘻哈和流行音乐方面则明显下滑。

等人，2025年），并且一致发现，模型仅处理人声音轨时，性能会有所提升。我们将探究，在基准测试涵盖的各项任务中，音频语言模型是否也能获得同样的益处。

我们使用 BS-RoFormer 模型（Lu 等人，2023年）将基准测试中的混合音频替换为人声音轨，并在相同配置下重新评估所有模型。如图 6所示，人声音轨会在不同程度上降低各项任务的表现。总体而言，模型在大多数任务上的表现都变差，这表明混合音频包含更丰富的信息，而且音源分离无法可靠地提升对音乐结构和语义的推理表现。

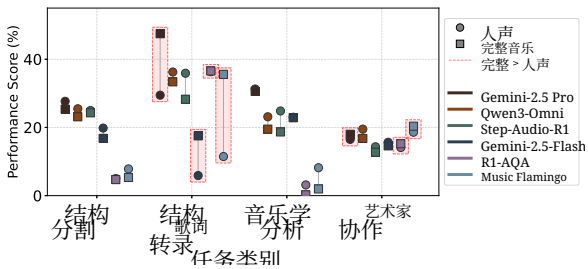


图 6：音源分离会降低音乐理解任务的表现。混合音频（方块）的表现始终优于人声音轨（圆圈），尤其是在结构化歌词转录任务中。

哪些音乐流派对音频语言模型构成的挑战最大？ BASS 收录了涵盖各种音乐流派的录音，如图 8所示。对于这项分析，我们聚焦于音乐学任务，并分析模型性能的平均值，以了解哪些音乐流派对音频语言模型最具挑战性。如图 5所示，模型在乡村音乐和摇滚上的表现相对较差。这表明，像流行音乐这样广受欢迎的音乐流派，可能在当前音频语言模型的训练数据中占比不足。

逐步推理能否提升音乐理解的表现？ BASS 中的许多任务都需要组合能力，可能会受益于逐步推理。虽然许多音频语言模型并未针对混合推理模式进行训练（即能够在更深入的逐步推理模式和更快速的直接响应模式之间切换），但我们研究了此类推理能力对基准测试中所有任务表现的影响。我们在 Qwen3-Omni 上进行了这项分析。[6] 在实验中，我们使用 vLLM 的推理控制方法[7]，在“思考”和“非思考”模式之间切换 Qwen3-Omni，并使用三个独立的随机种子进行所有实验。

我们的结果见 表3，表明启用“思考”后，BASS 的 12 项任务中有 6 项的模型性能显著提升。在艺人归属（+24.53）、基因归属（+17.97）和段落歌词转录（+13.33）上的性能提升幅度最大；经 Holm-Bonferroni 校正后，这些提升均具有统计显著性（$p < 0.001$）。完整歌词转录（-12.77）和基因优势排序（-2.11）等任务的性能略有下降，表明“思考”会降低这些任务的性能。

[6] Gemini Flash 等闭源模型也有“思考”和“非思考”版本，但目前尚不清楚它们是否使用相同的底层模型。

[7] https://docs.vllm.ai/en/latest/features/reasoning_outputs/#server-level-default-chat-template-kwargs

Task	无思考	思考	Δ
艺术家合作			
艺术家计数	25.04	24.59	-0.46
艺术家时长	4.59	8.97	+4.38
艺人归属	11.10	35.63	+24.53
艺人定位	3.64	7.29	+3.65
音乐学分析			
单基因检测	45.47	46.27	+0.80
成对基因检测	3.30	4.91	+1.61
基因优势排序	17.79	15.68	-2.11
基因归因	13.00	30.97	+17.97
结构分段			
完整结构分段	24.05	27.19	+3.14
乐段结构分割	15.06	19.84	+4.77
结构化歌词转录			
完整结构化歌词转录	28.08	15.31	-12.77
段落结构化歌词转录	27.86	41.19	+13.33

表3：Qwen3-Omni 启用和未启用“思考”时的性能比较。启用“思考”后取得统计显著提升的任务（在三个随机种子上进行 Holm-Bonferroni 校正后，$p < 0.001$）以绿色突出显示。

音频语言模型是否难以将记忆中的元数据与音乐音频建立联系？ BASS 包含涵盖多种音乐流派的音乐录音，尤其是流行歌曲。鉴于许多近期的语言模型都基于大规模网络数据进行训练，而这些数据可能包含重叠的音乐内容或元数据，因此，有必要评估模型性能在多大程度上反映的是记忆能力，而非基于音频的推理能力。为此，我们仅向模型提供艺人名称和歌曲名称以及问题，而不提供音频。随后，我们将这种纯文本条件下的性能与提供音乐录音时的性能进行比较。本分析在 Qwen3-Omni、Gemini 和 Gemini Pro 上进行，重点关注艺术家合作、音乐学分析和结构化歌词转录任务。由于这些评估不提供任何音频，我们排除了需要时间戳定位的问题。

我们的结果见图7。结果表明，与仅提供音频相比，仅提供元数据（艺人名称和歌曲名称）竟能显著提升歌词转录任务的表现，其中 Gemini Flash 的提升幅度最大。尽管这一提升说明模型能够利用训练数据中记忆的艺人与歌曲关联，但也凸显了模型在缺少文本先验时面临的挑战，表明仍有很大的改进空间。在其他任务中，与音频条件相比，表现要么下降，要么基本不变。这符合我们的预期，因为这些任务更多依赖音频中包含的信息，而非音频，而不是记忆中的元数据。尽管音频语言模型的语言模型组件可能在训练期间见过这些歌词，但音频语言模型在推理时仍难以将这些先验知识与音乐音频联系起来，这仍有些令人意外。

5 相关工作

现有包含音乐相关任务的基准大多是广泛的音频基准，旨在评估大型音频语言模型的通用音频理解能力。示例包括 BLAB（Ahia 等人，2025年）、MMAU（Sakshi 等人，2025年）、MMAU-Pro（Kumar 等人，2025年）、MMAR（Ma 等人，2025b）以及 AIR-Bench（Yang 等人，2024年）。其中一些基准将音乐视为众多音频领域之一，评估主要侧重于一般声学理解，而非音乐特定的推理。MuChoMusic（Weck等人，2024年）则专门关注音乐理解。该基准使用 MusicCap 和歌曲描述数据集（SDD）（Manco 等人，2023年）构建多项选择题。然而，MusicCaps 仅限于 10 秒的片段，SDD 则仅限于两分钟的音频片段，因此无法评估长距离音乐结构。CMI-Bench（Ma 等人，2025a）确实包含开放式任务，但其核心是经典的音乐信息检索（MIR）任务，并且是将传统的音乐信息检索标注重新表述为指令跟随格式，而不是进行长篇音乐推理。MusicQA（Liu 等人，2023年）作为 Music Understanding LLaMA 工作的一部分被提出，为音乐提供问答和描述生成能力，但其任务多样性局限于一般音乐知识和生成，而非在延展的音乐上下文中进行全面推理。我们的 BASS 专门用于评估在延展时间上下文中的语义和结构音乐理解与推理能力。

6 结论

在本文中，我们介绍了 BASS，旨在评估人工智能系统中的音乐理解与推理能力。BASS 涵盖四个类别——结构分段、歌词转录、音乐学分析和艺术家合作 ——共包含 12 项任务，要求同时进行结构性和语义推理。我们评估了 14 个开源及前沿音频语言模型，其中包括采用混合思考训练的模型，并发现即使是表现最佳的模型，其平均表现也仅达到 26%。我们的分

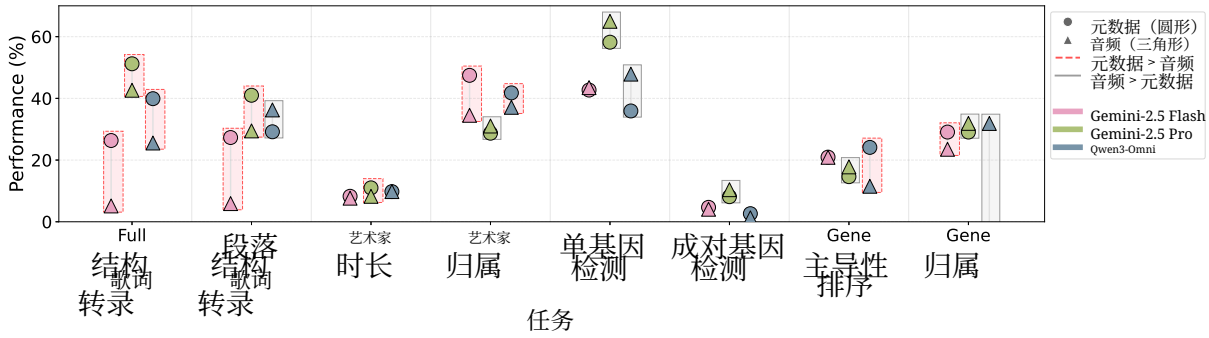


图7：将音乐音频替换为元数据（艺人名称和歌曲名称）时的性能对比。仅使用元数据时，模型在歌词转录上的相对性能高于其他任务，这表明模型在更大程度上依赖于训练数据中记忆的艺人与歌曲关联。

析表明，最先进的模型在结构分段和艺术家合作等更高层次的推理任务上表现吃力，而在歌词转录方面相对更好。BASS 突出了模型的优势与不足，为未来开发具备强大长篇音乐理解和推理能力的音频语言模型提供指导。

7 局限性

BASS 中的歌曲主要为英语歌曲，或由英语艺人演唱。我们承认，本研究的范围聚焦于英语内容，因为许多现有的音频大语言模型主要基于英语数据集进行训练。这些模型在该基准测试上的表现可能无法完全推广到多语言或多文化环境中的音乐。我们将这一探索留待未来工作。我们的数据整理流程借鉴了现有的音乐资源。为确保质量，我们进行了严格筛选，以确保所有数据均来自可信来源，并由专业音乐学家进行标注。

伦理声明

BASS 的音乐资源来源于采用 Creative Commons 许可的作品，音频取自 YouTube。由于许可限制，我们计划不直接分发音乐音频文件。相反，发布数据集时，我们将提供每首曲目对应的 YouTube 网址链接，使用户能够在尊重版权的同时合法访问音频。尽管我们整理数据集时重点关注公开可用的内容，但用户应注意其中包含多位不同艺术家的作品；我们鼓励大家负责任、尊重地使用这些作品。为提高透明度和可复现性，我们将公开基准测试以及相关评估指标和数据整理框架，供研究社区使用，以便大家为我们的工作做出贡献并在此基础上进一步研究。

参考文献

Andrea Agostinelli、Timo I. Denk、Zalán Borsos、Jesse Engel、Mauro Verzetti、Antoine Caillon、Qingqing Huang、Aren Jansen、AdamRoberts、Marco Tagliasacchi、Matt Sharifi、Neil Zeghidour和Chris- tian Frank。2023年。MusicLM：从文本生成音乐。预印本，arXiv:2301.11325。

Vanshaj Agrawal和Sunil Karamchandani。2023年。应用于人声-伴奏提取的音频源分离。*e-Prime*——电气工程、电子与能源进展，5:100254。

Orevaoghene Ahia、Martijn Bartelds、Kabir Ahuja、Hila Gonen、Valentin Hofmann、Siddhant Arora、Shuyue Stella Li、Vishal Puttagunta、Mofetoluwa Adeyemi、Charishma Buchireddy、Ben Walls、Noah Bennett、Shinji Watanabe、Noah A. Smith、Yulia Tsvetkov和Sachin Ku mar。2025年。BLAB：极长音频基准。预印本，arXiv:2505.03054。

Michael Castelluccio. 2006年. 音乐基因组计划。*Strategic Finance*。

Keunwoo Choi、George Fazekas 和 Mark Sandler。2016年。使用基于曲目内转移训练的循环神经网络生成播放列表的算法。预印本，arXiv:1606.02096。

Ond˘rej Cífka、Hendrik Schreiber、Luke Miner 和 Fabian-Robert Stöter。2024年。《面向人类的歌词转录：可读性感知基准》。即将发表；预印本 arXiv:2408.06370。

Jacopo de Berardinis、Angelo Cangelosi 和 Eduardo Coutinho。2020年。《音乐情感的多重声音：通过音源分离提升音乐情感识别模型及其可解释性》。载于

第 *21* 届国际音乐信息检索协会会议（*ISMIR*）论文集，第 310–317 页，加拿大蒙特利尔。Soham Deshmukh、Satvik Dixit、Rita Singh 和 Bhiksha Raj。2025年。Mellow：用于推理的小型音频语言模型。预印本，arXiv:2503.08540。Michèle Duguay、Kate Mancey 和 Johanna Devaney。2023年。合作歌曲数据集（CoSoD）：流行音乐中多艺术家合作的标注数据集。载于第 *24* 届国际音乐信息检索协会会议（*ISMIR 2023*）论文集，第 71–79 页。

Michèle Duguay. 2022. 分析录音虚拟空间中的人声摆位。*Music Theory Online*，28(4):1–?

Ben Duinker. 2021. 歌曲形式与嘻哈音乐的主流化. 107:93–135.

Andreas F Ehmann、Mert Bay、J Stephen Downie、Ichiro Fujinaga和David De Roure。2011。音乐结构分割算法评估：对MIREX 2010分析与数据集的扩展。见第*12*届国际音乐信息检索学会会议，*ISMIR 2011*，第561–566页。

Gemini Team、Rohan Anil、Jean-Baptiste Alayrac、Jiahui Yu、Radu Soricut、Johan Schalkwyk等……2023年。Gemini：一系列能力强大的多模态模型。*arXiv*预印本 *arXiv:2312.11805*。

Gemma Team、Aishwarya Kamath、Johan Ferret、Shreya Pathak、Nino Vieillard、RamonaMerhej、Sarah Perrin、Tatiana Matejovicova、Alexandre Ramé、Morgane Rivière、Louis Rouillard、Thomas Mesnard、Geoffrey Cideron、J ean bastien Grill、Sabela Ramos、Edouard Yvinec、Michelle Casbon、Etienne Pot、Ivo Penchev及另外197位作者。2025年 Gemma 3 技术报告。预印本，arXiv:2503.19786。

Sreyan Ghosh、Arushi Goel、Lasha Koroshinadze、Sang gil Lee、Zhifeng Kong、Joao Felipe Santos、Ramani Duraiswami、Dinesh Manocha、Wei Ping、Moham- mad Shoeybi 和 Bryan Catanzaro。2025a。Music flamingo：扩展音频语言模型中的音乐理解。预印本，arXiv:2511.10289。Sreyan Ghosh、Zhifeng Kong、Sonal Kumar、S Sak- shi、Jaehyeon Kim、Wei Ping、Rafael Valle、Dinesh Manocha 和 Bryan Catanzaro。2025b。Audio flamingo 2：具备长音频理解和专家级推理能力的音频语言模型。发表于第四十二届国际机器学习会议。Arushi Goel、Sreyan Ghosh、Jaehyeon Kim、Sonal Ku- mar、Zhifeng Kong、Sang-gil Lee、Chao-Han Huck Yang、Ramani Duraiswami、Dinesh Manocha、Rafael Valle 和 Bryan Catanzaro。2025。Audio flamingo 3：以完全开放的大型音频语言模型推进音频智能。*arXiv* 预印本 *arXiv*。

Chien-yu Huang、Wei-Chih Chen、Shu-wen Yang、Andy T. Liu、Chen-An Li、Yu-Xiang Lin、Wei-Cheng Tseng、Anuj Diwan、Yi-Jen Shih、Jiatong Shi、William Chen、Chih-Kai Yang、Xuanjun Chen、Chi-Yuan Hsiao、Puyuan Peng、Shih-Heng Wang、Chun- Yi Kuan、Ke-Han Lu、Kai-Wei Chang 等另外59位作者。2025。Dynamic-SUPERB 第二阶段：一个通过180项任务测量口语语言模型能力、以协作方式不断扩展的基准测试。发表于国际学习表征会议论文集（*ICLR*）。

KimiTeam、Ding Ding、Zeqian Ju、Yichong Leng、Songxiang Liu、Tong Liu、Zeyu Shang、Kai Shen、Wei Song、Xu Tan、Heyi Tang、Zhengtao Wang、Chu Wei、Yifei Xin、Xinran Xu、Jianwei Yu、Yu- tao Zhang、Xinyu Zhou、Y. Charles及另外21位作者。2025年。Kimi-Audio 技术报告。*arXiv* 预印本 *arXiv:2504.18425*。

RobertKomaniecki. 2017. 分析说唱音乐中的协作流动。*Music Theory Online*，23(3)。

Sonal Kumar、Šimon Sedlá˘cek、Vaibhavi Lokegaonkar、Fernando López、Wenyi Yu、Nishit Anand、Hyeong- gon Ryu、Lichang Chen、Maxim Pli˘cka、Miroslav Hlavá˘cek 等。2025年。MMAU-Pro：一项用于全面评估音频通用智能的高难度综合基准测试。*arXiv* 预印本 *arXiv:2508.13992*。

Gang Li、Jizhong Liu、Heinrich Dinkel、Yadong Niu、Junbo Zhang 和 Jian Luan。2025年。强化学习优于监督微调：音频问答案例研究。预印本，arXiv:2503.11197。

Alexander H. Liu、Andy Ehrenberg、Andy Lo、Clément Denoix、Corentin Barreau、Guillaume Lample、Jean-Malo Delignon、Khyathi Raghavi Chandu、Patrick vonPlaten、Pavankumar ReddyMuddireddy、Sanchit Gandhi、Soham Ghosh、Srijan Mishra、Thomas Fou- bert、Abhinav Rastogi、Adam Yang、Albert Q. Jiang、Alexandre Sablayrolles、Amélie Héliou 等 87 人。2025年。Voxtral。预印本，arXiv:2507.13264。

Shansong Liu、Atin Sakkeer Hussain、Chenshuo Sun 和 Ying Shan。2023年。Music understanding llama：通过问答和描述生成推进文本到音乐生成。预印本，arXiv:2308.11276。Wei-Tsung Lu、Ju-Chiang Wang、Qiuqiang Kong 和 Yun-Ning Hung。2023年。使用 Band-Split RoPE Transformer 进行音乐源分离。预印本，arXiv:2309.02612。Yinghao Ma、Siyou Li、Juntao Yu、Emmanouil Benetos 和 Akira Maezawa。2025年a。Cmi-bench：用于评估音乐指令遵循的综合基准测试。预印本，arXiv:2506.12285。Ziyang Ma、Yinghao Ma、Yanqiao Zhu、Chen Yang、Yi-Wen Chao、Ruiyang Xu、Wenxi Chen、Yuanzhe Chen、Zhuo Chen、Jian Cong、Kai Li、Keliang Li、Siyou Li、Xinfeng Li、Xiquan Li、Zheng Lian、Yuzhe Liang,

Minghao Liu、Zhikang Niu 和另外 15 位作者。2025年b。MMAR：语音、音频、音乐及其混合内容深度推理的挑战性基准测试。载于第三十九届神经信息处理系统年会数据集与基准赛道。

Ilaria Manco、Benno Weck、Seungheon Doh、Minz Won、Yixiao Zhang、Dmitry Bogdanov、Yusong Wu、Ke Chen、Philip Tovstogan、Emmanouil Benetos、Elio Quinton、György Fazekas 和 Juhan Nam。2023年。歌曲描述数据集：用于音乐与语言评估的音频描述语料库。载于*NeurIPS 2023* 音频机器学习研讨会。

Annamaria Mesaros 和 Tuomas Virtanen。2008年。歌唱歌词的自动识别。载于第*11*届数字音频效果国际会议（*DAFx*）论文集，第1–4页。

微软、：、Abdelrahman Abouelenin、Atabak Ashfaq、Adam Atkinson、Hany Awadalla、Nguyen Bach、Jianmin Bao、Alon Benhaim、Martin Cai、Vishrav Chaudhary、Congcong Chen、Dong Chen、Dong- dong Chen、JunkunChen、Weizhu Chen、Yen-Chun Chen、Yi ling Chen、Qi Dai 等57人。2025年。Phi-4-mini 技术报告：通过 LoRA 混合实现紧凑而强大的多模态语言模型。预印本，arXiv:2503.01743。

Nicola Montecchio、Pierre Roy 和 François Pachet。2020 年。音乐流媒体服务用户的跳过行为及其与音乐结构的关系。*PLOSONE*，15(9)：e0239418。

Music Tomorrow。2025 年。Spotify 推荐系统内部：完整指南（2025 年更新）。访问日期：2026-01-04。介绍 Spotify 如何使用时间结构单元进行基于内容的推荐。

Oriol Nieto、Matthew C. McCallum、Matthew E. P. Davies、Andrew Robertson、Adam M. Stark 和 Eran Egozy。2019 年。Harmonix Set：西方流行音乐的节拍、强拍和功能段落标注。载于 第 *20* 届国际音乐信息检索学会会议论文集，*ISMIR 2019*，代尔夫特，荷兰，*2019* 年 *11* 月 *4* 日至 *8* 日，第 565–572 页。

Sebastian Oramas、Fabien Gouyon、Stuart Hogan、Chris Landau 和 Anahid Ehmann。2025 年。Mgphot：一个流行音乐（1958年至2022年）音乐学标注数据集。《国际音乐信息检索学会汇刊》，8（1）：108–120。

Longshen Ou、Xiangming Gu 和 Ye Wang。2022 年。将 wav2vec 2.0 的迁移学习用于自动歌词转录。载于第 *23* 届国际音乐信息检索学会会议（*ISMIR*）论文集，第 891–899 页，印度班加罗尔。

Jouni Paulus、Meinard Müller 和 AnssiKlapuri。2010年。基于音频的音乐结构分析领域现状报告分析。载于 第十一届国际音乐信息检索学会会议论文集，*ISMIR 2010*，荷兰乌得勒支，*2010*年*8*月*9*日至*13*日，第625–636页。国际音乐信息检索学会。Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。2023年。通过大规模弱监督实现稳健的语音识别。载于 第四十届国际机器学习会议论文集，ICML'23。J MLR.org。Aldona Rosner 和 Bozena Kostek。2018年。基于乐器音轨分离的自动音乐流派分类。《智能信息系统杂志》，50(2)：363–384。

S Sakshi、Utkarsh Tyagi、Sonal Kumar、Ashish Seth、Ramaneswaran Selvakumar、Oriol Nieto、Ramani Duraiswami、Sreyan Ghosh 和 Dinesh Manocha。2025 年。MMAU：大规模多任务音频理解与推理基准。载于第十三届国际学习表征会议。

Jaza Syed、Ivan Meresman Higgs、Ondˇrej Cífka 和 Mark Sandler。2025年。利用音乐源分离通过 Whisper 进行自动歌词转录。预印本，arXiv:2506.15514。

Changli Tang、Wenyi Yu、Guangzhi Sun、Xianzhao Chen、Tian Tan、Wei Li、Lu Lu、Zejun Ma 和 Chao Zhang。2024年。Salmonn：面向大语言模型的通用听觉能力。预印本，arXiv:2310.13289。

Fei Tian、Xiangyu Tony Zhang、Yuxin Zhang、Haoyang Zhang、Yuxin Li、Daijiao Liu、Yayue Deng、Dong- hang Wu、Jun Chen、Liang Zhao、Chengyuan Yao、Hexin Liu、Eng Siong Chng、Xuerui Yang、Xiangyu Zhang、Daxin Ji ang 和 Gang Yu。2025年。Step-audio-r1 技术报告。预印本，arXiv:2511.15848。

Ju-ChiangWang、Yun-Ning Hung 和 Jordan BLSmith。2022年。捕捉副歌、主歌、前奏或其他部分：使用结构功能分析一首歌曲。载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 416–420 页。IEEE。

Jun-You Wang、Chon-In Leong、Yu-Chen Lin、Li Su 和 Jyh-Shing Roger Jang。2023年。调整预训练语音模型以进行普通话歌词转录和对齐。载于 *2023 IEEE* 自动语音识别与理解研讨会（*ASRU*），第 1–8 页。IEEE。

Benno Weck、Ilaria Manco、Emmanouil Benetos、Elio Quinton、George Fazekas 和 Dmitry Bogdanov。2024年。MuChoMusic：评估多模态音频语言模型中的音乐理解能力。载于第 *25* 届国际音乐信息检索学会会议（*ISMIR*）论文集。

Gijs Wijngaard、Elia Formisano、Michele Esposito 和 Michel Dumontier。2025年。Audsemthinker：通过对声音语义进行推理来增强音频语言模型。预印本，arXiv:2505.14142。

Zhifei Xie、MingbaoLin、Zihang Liu、Pengcheng Wu、Shuicheng Yan 和 Chunyan Miao。2025年。Audio-Reasoner：提升大型音频语言模型的推理能力。预印本，arXiv:2503.02318。

Jieping Xu、Xirong Li、Yun Hao 和 Gang Yang。2014年。音源分离提升音乐情绪识别能力。载于国际多媒体检索会议论文集，第 423–426 页。ACM。

Jin Xu、Zhifang Guo、Hangrui Hu、Yunfei Chu、Xiong Wang、Jinzheng He、Yuxuan Wang、Xian Shi、Ting He、Xinfa Zhu、Yuanjun Lv、Yongqi Wang、Dake Guo、He Wang、Linhan Ma、Pei Zhang、Xinyu Zhang、Hongkun Hao、Zishan Guo 等另外 19 位作者。2025年。Qwen3-omni：一种用于文本、图像、音频和视频的单一多模态模型。*arXiv* 预印本 *arXiv:2509.17765*。

Qian Yang、Jin Xu、Wenrui Liu、Yunfei Chu、Ziyue Jiang、XiaohuanZhou、Yichong Leng、Yuanjun Lv、Zhou Zhao、Chang Zhou 和 Jingren Zhou。2024年。AIR-Bench：通过生成式理解对大型音频语言模型进行基准测试。预印本，arXiv:2402.07729。

Le Zhuo、Ruibin Yuan、Jiahao Pan、Yinghao Ma、Yizhi Li、Ge Zhang、Si Liu、Roger Dannenberg、Jie Fu、Chenghua Lin、Emmanouil Benetos、Wenhu Chen、Wei Xue 和 Yike Guo。2023年。Lyricwhiz：通过向 ChatGPT 轻声提示，实现稳健的多语言零样本歌词转录。载于第*24*届国际音乐信息检索学会会议（*ISMIR*）论文集，意大利米兰。

附录

A 数据集详情

A.1 数据集统计

我们在 BASS 的各项任务中全面展示了数据统计，见表4。在图8中，我们展示了数据集中音乐流派的分布，其中流行音乐（30.3%）和嘻哈（25.6%）占比最高，其次是电子音乐、摇滚和节奏布鲁斯等多种类别。

Cat.	Task	#Q	# S	平均值	最小值	最大值
结构	FSS	74	74	3.06	1.70	10.78
	SSS	224	224	3.47	1.93	9.31
	总体	298	298	3.37	1.70	10.78
歌词	FSLT	665	665	3.73	1.60	11.72
	SSLT	170	170	3.97	2.17	8.44
	总体	884	884	3.78	1.60	11.72
音乐。	SGD	335	335	3.15	1.86	9.23
	PGD	147	147	3.77	0.82	6.37
	GDR	33	33	3.60	2.58	4.95
	GA	94	94	14.14	9.77	19.44
	总体	609	484	5.57	0.82	19.44
合作	Cnt	229	229	3.92	2.37	6.97
	Dur	327	327	3.83	1.34	6.97
	Loc	247	247	3.83	2.37	6.13
	Attr	64	64	3.88	1.34	6.97
	总体	867	327	3.83	1.34	6.97

表4：各项任务的数据统计。

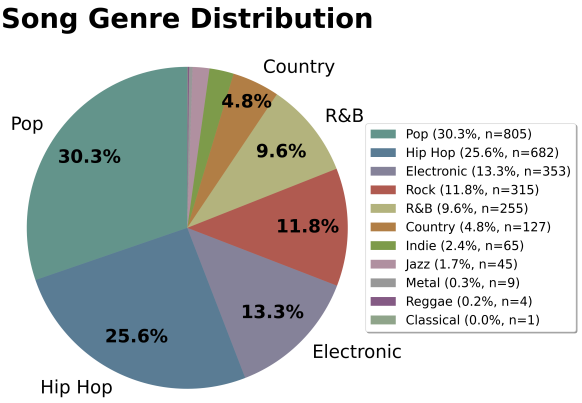


图8：BASS 中各首歌曲的音乐流派分布

A.2 所有任务的提示词

在表5中，我们列出了基准测试中所有任务的具体提示词和回答格式说明。

B 特定任务结果

在§4 中，我们在论文正文中概述了主要发现，并讨论这些发现对音频语言模型在音乐结构、歌词、艺人角色和音乐流派方面的推理能力有何启示。此处，我们将展示各项任务的细粒度结果。

B.1 结构分段

对于结构分段任务，我们的结果表明，模型在定位一首歌曲中的所有段落（完整分割）时，通常比定位单个孤立段落（分段级分段）表现更好。在所有评估模型中，这两项任务之间的平均绝对归一化分数差异为7.63%。在图9中，我们展示了表现最佳模型在各段落上的平均模型性能。我们观察到，模型更擅长分割歌曲中较早出现的段落，如前奏和主歌（第一主歌）。随着歌曲推进至尾奏部分，性能逐渐下降。

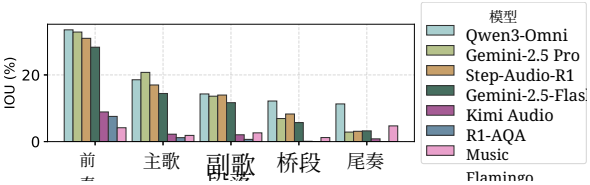


图9：结构分段的各段落性能。较早段落的性能最高，并随着录音接近结尾而逐渐下降。

图13展示了所有模型的混淆矩阵。我们发现，主歌和副歌最容易被彼此混淆。我们并不认为仅凭声学信息就能区分不同歌曲中的主歌和副歌；副歌中重复的旋律和歌词等线索应该更有帮助。我们推测，这种混淆是因为音频语言模型并未经过专门训练，无法理解跨长程依赖的音乐结构。

B.2 结构化歌词转录

对于这项任务，我们观察到的总体趋势与结构分段一致：模型在完整歌曲转录上的表现显著优于分段级转录。在所有评估的模型中，完整歌曲转录与分段级转录之间的平均归一化表现差异为 4.35%（通过 IWER §3测得）。

虽然整体表现模式与结构分段中的情况相似，但我们注意到前奏段落存在明显差距。我们推测，这是因为前奏和尾奏通常没有歌词；因此，模型更容易产生插入错误，可能会幻觉出词语，或重复其他段落中的短语。

为探究这一点，我们分析了细粒度词错误率和误差类型。我们发现，模型在主歌中产生的删除更多，而在前奏和尾奏（附录，图12）中产生的插入更多，导致这些段落的转录表现较差。我们进一步通过量化预测转录文本中的重复短语来刻画这一行为，使用

$$\frac{\sum_{\phi \in \text{repeats}} \max(0, c_\phi - r_\phi) \cdot |\phi|}{|\text{ref_words}| + k},$$

其中 ϕ 是重复短语，c_ϕ 是该短语在预测结果中的数量，r_ϕ 是该短语在参考文本中的数量，$|\phi|$ 是短语的词数，$|\text{参考_词}|$ 是参考文本中的总词数，而 k 是平滑常数，用于避免较短参考文本的得分虚高。如图13所示，重复错误确实最常出现在前奏和尾奏段落中，无论是完整转录还是分段级转录都是如此，这支持了我们的假设。

B.3 音乐学分析

在这项任务中，我们观察到模型在单基因检测任务上的表现最佳：模型需要从四个选项中选出一项在歌曲中占主导地位的音乐学属性（例如，是否存在节奏律动）。在双基因检测任务中，表现大幅下滑；模型必须从八个选项中选出一对属性，因此任务更具挑战性；即使是 Gemini，准确率也只有约 10%。基因优势排序的表现同样不佳，该任务要求按主导程度对属性进行排序。不过，在基因归因任务中，结果有所改善：模型需要从四首歌曲中选出一首具有主导属性的歌曲，任务层级更高。

B.4 艺术家合作

在所有任务类别中，这一类别似乎最具挑战性。尽管总体表现较低，但模型在艺术家计数和艺人归属任务上的表现始终优于其他任务。这些任务都侧重于识别演唱风格、艺人角色和时间定

位。以往的基准测试报告指出，音频语言模型缺乏时间定位能力（Ahia 等人，2025年；Kumar 等人，2025年；Ma 等人，2025b），因此这一结果并不令人意外。在图 10中，我们分析了需要以不同粒度识别歌曲中艺人角色的任务。模型通常更准确地识别主艺人，而在艺人既非主艺人也非特邀艺人的情况下表现最差。

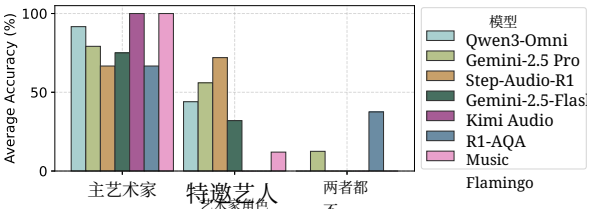


图 10：模型在不同艺人角色中的准确率。主艺人的准确率最高，特邀艺人较低，不属于这两类的艺人准确率最低。

对于演唱风格呈现，我们分析了识别歌唱和说唱（图 4）方面的表演差异。大多数模型在说唱相关问题上的准确率更高，其中 Gemini Flash 的差距最大。

最后，我们量化了艺术家计数和时长相关任务中的预测误差（见附录 图 14 和15）。Gemini Pro 的误差接近于零，而 Audio-Flamingo、Audio-Reasoner 和 Phi-mm Instruct 倾向于低估时长，Gemini Flash 和 R1 AQA 则倾向于高估。艺术家计数中也出现了类似趋势，大多数模型都低估了完整歌曲中的艺术家数量。由于 Gemma、Qwen 等模型产生的幻觉误差超过 10^7 秒，且会主导图表的刻度，因此未将这些极端离群值纳入图中。

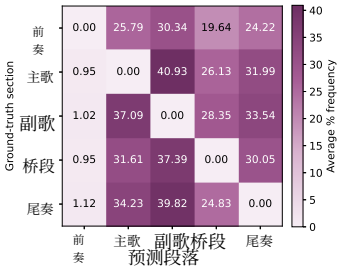


图 11：各模型在结构分段任务中的平均段落预测误差，突显了副歌与主歌乐段之间频繁出现的混淆。

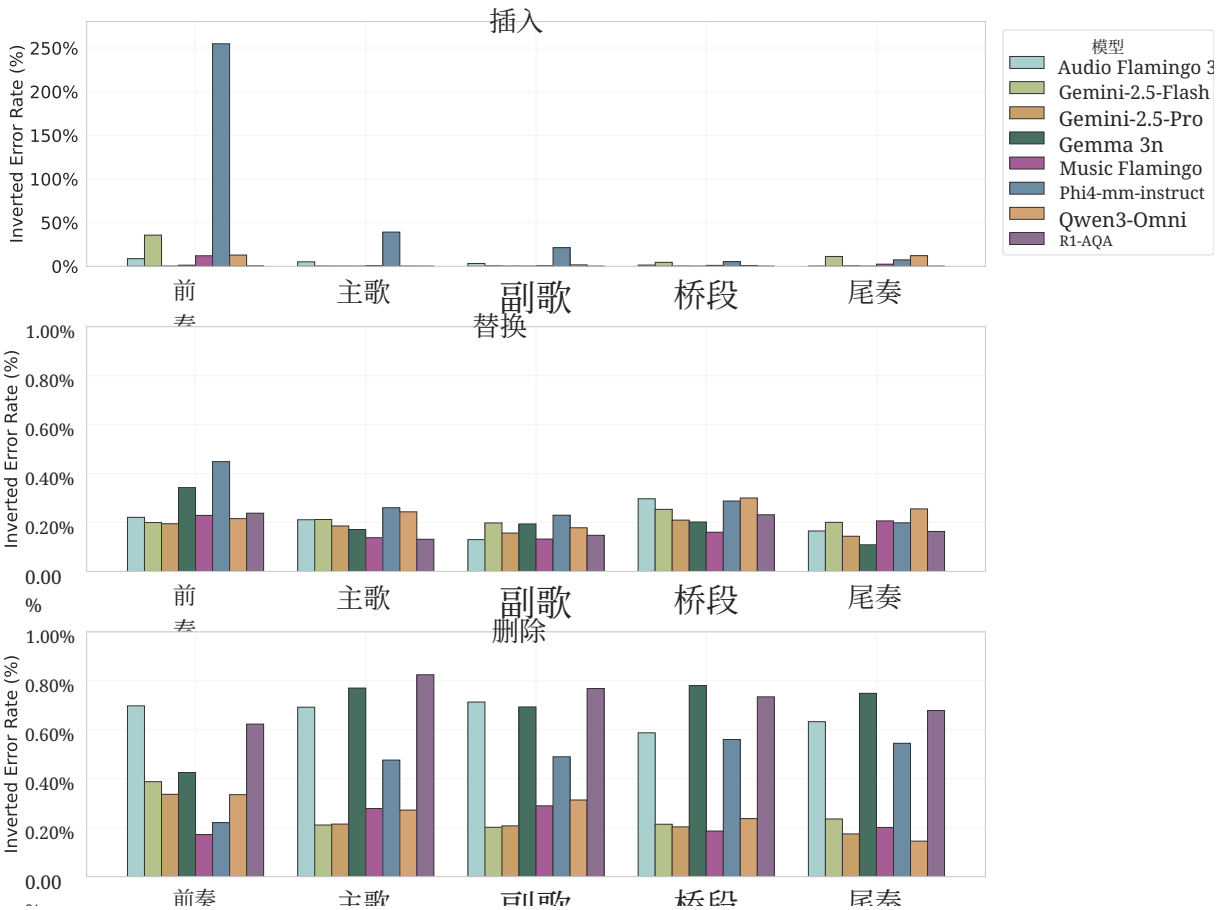


图12：按歌曲段落分析反转错误率。各段落中反转**插入**、**替换**和**删除** 错误率的细分情况。请注意，与**替换**和**删除**（低于 1.00%）相比，**插入** 错误的比例显著更高**(最高达250%)**，尤其是在前奏 段落中。

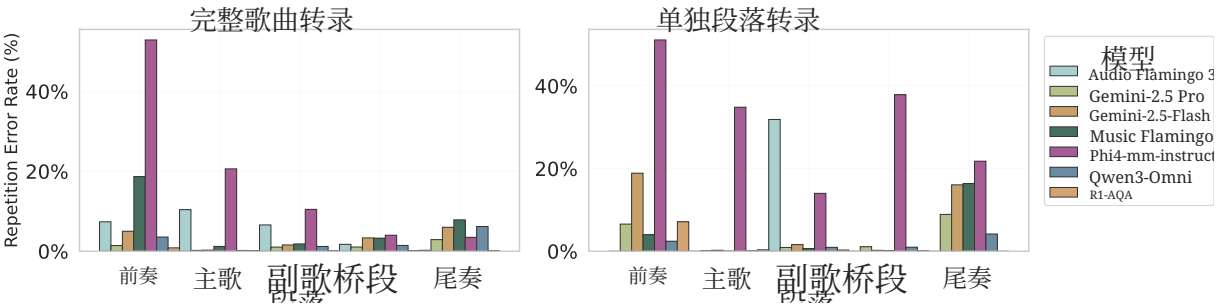


图13：完整转录和分段转录各段落的重复错误率。在前奏 中，两项任务都出现了较高的重复错误（短语幻觉）；分段转录的重复率始终高于完整转录。

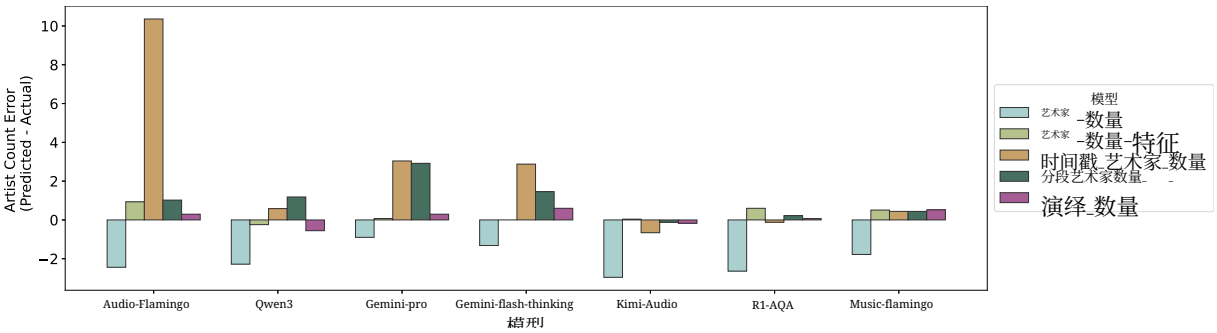


图 14：艺术家计数子任务的模型性能。y 轴显示误差（预测值 – 实际值），表示艺术家和配送数量的高估（正值）或低估（负值）。

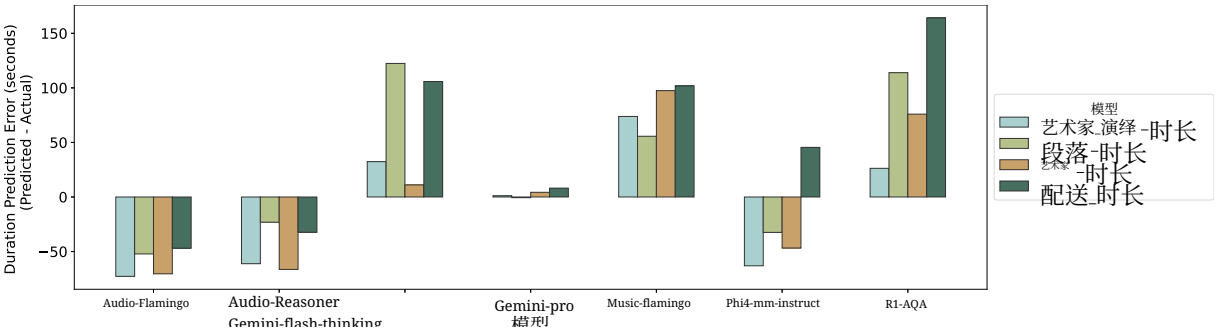


图 15：艺术家时长任务中的时间预测误差（秒）。y 轴显示误差（预测值 – 实际值），表示时长的高估（正值）或低估（负值）。

类别	Task	子任务	示例问题与预期输出格式
结构分割	完整结构分段-tion	–	将曲目划分为对应于以下任一类别的段落：主歌、副歌、桥段、前奏、尾奏、预副歌、后副歌。对于每个段落，请提供以下信息：段落标题、整个段落的起始和结束时间戳。**格式：** [{"section": "<标题>" , "开始": <时间戳>, "结束": <ts>} , . . .]
	段落结构分段	–	给定歌曲，请提供所有主歌乐段的开始和结束时间戳（以秒和毫秒为单位）所有主歌乐段的。**格式：** [{"start": <ts> , "结束": <ts> }, ...]
歌词转录	Full 结构 歌词转录	–	已提供一首歌曲的结构分段。请转录每个段落的歌词。**格式：** [{"section": "<段落1>", "歌词": "....." }, ...]
	段落结构化歌词转录	–	你能只提供所有主歌乐段的歌词吗？每个段落的转录——都应作为一个独立项目处理，一个列表。**格式：** [{"段落": "段落_名称", "歌词": "..." }]
音乐学分析	单基因检测	–	根据以下音乐属性，哪一项在这首歌中最为突出？选项：复合拍子、打击乐、人声粗粝度、合成器。**格式：** 仅回答音乐学特征的名称。
	成对基因检测	–	这首曲目中最突出的两个特征是什么？**格式：** 仅以两个名称的列表作答。示例：['<属性1>', '<属性2>']
	基因归因	–	哪首歌曲的人声音域最高？（该属性衡量主唱的音高范围的主唱）。**格式：** 仅回复与该歌曲对应的数字（1、2、3 或 4）。
	基因主导性排名-ing	–	将以下节奏模式和律动按从最不突出到最突出的顺序排列。**格式：** 仅回复按名次排列的名称列表。示例：['<属性1>', '<属性2>']
艺术家协作	艺术家计数	标准 / 特邀	有多少位艺术家参与演出？ / 有多少位表演者是特邀艺人艺人？
		演唱方式	这首歌曲中有多少位艺人在演唱说唱段落？
		段落 / 时间	后副歌中有多少位艺人在演唱？ / 从 211 秒到 249 秒？**格式（所有计数）：** 仅回答数字。
	艺术家时长	目标 / 人声	艺人2的表演总时长是多少？ / 有多少秒是用于说唱的？
		艺人演绎 / 段落	艺人1说唱部分的总时长是多少？ / 桥段持续多长时间？**格式（全部时长）：** 仅以秒为单位作答，精确到毫秒，包括毫秒。
	艺人定位	–	艺人2在这首歌曲中从第几秒开始演唱？**格式：** 仅以秒为单位并精确到毫秒，回答时间戳数字。
	艺人归属	人声对比	第一段主歌中的特邀艺人与第二段主歌中的主艺术家是否使用相同类型的演唱方式？**格式：** 仅回答：相同或不同。
		段落风格/角色	对于第一个副歌，艺术家是在说唱、歌唱，还是两者都不是？是主艺术家还是特邀艺人？**格式：** 说唱、歌唱、两者皆非，或主艺术家、特邀艺人、两者都不是。

表5：所有类别的任务、示例问题和具体格式说明。

		音乐学分析 结构分段						艺术家合作				歌词转录	
模型	**Size**	SGD 投影梯度下降 GDR			GA	Full	段落	Cnt.	Dur.	Loc.	属性	Full	段落
Gemini-2.5 Pro	-	73.73	10.88	21.20	48.90	32.34	22.98	25.77	8.25	3.12	56.66	1.35	2.40
Qwen3-Omni	30B	60.90	6。16	15。20	48。90	28。88	22.07	18.54	9.78	9.38	63.66	2.92	1.76
Step-Audio-R1	-	57.91	10.27	18.18	48.94	32.19	17.73	19.34	6.42	7.81	53.13	3.57	3.61
Gemini-2.5-Flash	-	57.61	6.82	24.20	42.60	24.47	15.15	20.40	7.66	3.12	56.81	18.51	15.97
Music Flamingo	8B	49.25	4.76	0.00	25.50	13.07	2.59	33.54	4.27	4.69	56.97	5.36	7.72
R1-AQA	8.2B	34.03	0.00	0。00	30。90	5.91	4.16	35.03	0.92	0.00	50.84	1.50	1.75
Gemma 3	5B	35.22	4.76	3.03	31.91	8.92	4.35	12.25	0.30	1.56	40.80	1.37	1.70
Audio Flamingo 3	-	44.20	8.18	0。00	30。90	7.03	0.01	21.92	1.84	0。00	49。80	4.68	2.76
Voxtral-Small-24B	24B	37.61	8.84	9.12	17.00	32.16	8.84	9.09	2.76	0.00	38.52	2.78	15.46
SALMONN-7B	7B	25.37	1.36	0.30	28.70	3.18	0.12	27.32	0.61	0.00	34.86	7.61	20.59
Kimi Audio	7B	37.61	4.76	0.00	0。00	13。06	2.64	19.12	0.92	0.00	52.07	18.33	54.56
Phi4-mm-instruct	5.5B	33.73	1.36	9.12	21.30	2.71	0.16	23.59	3.67	0。00	38。51	79。29	364.54
Mellow	167M	0.30	0.00	0.00	0.00	0.00	0.00	0.00	2.14	0.00	0.00	19.02	3.46
Audio-Reasoner	-	33.73	3.40	6.06	30.85	6.63	2.91	18.47	3.06	0.00	2.31	-	-

表6：经评估的模型在 BASS 各项任务中的原始性能得分（**归一化前**），涵盖音乐学分析（单基因检测—SGD、成对基因检测—PGD、基因优势排序—GDR 和歌曲级基因检测—Song-GD）、结构分段、艺术家合作和歌词转录。音乐学分析和艺术家合作任务的指标为完全匹配准确率（%）。结构分段的指标为交并比（%）。歌词转录的指标为词错误率（非百分比）。

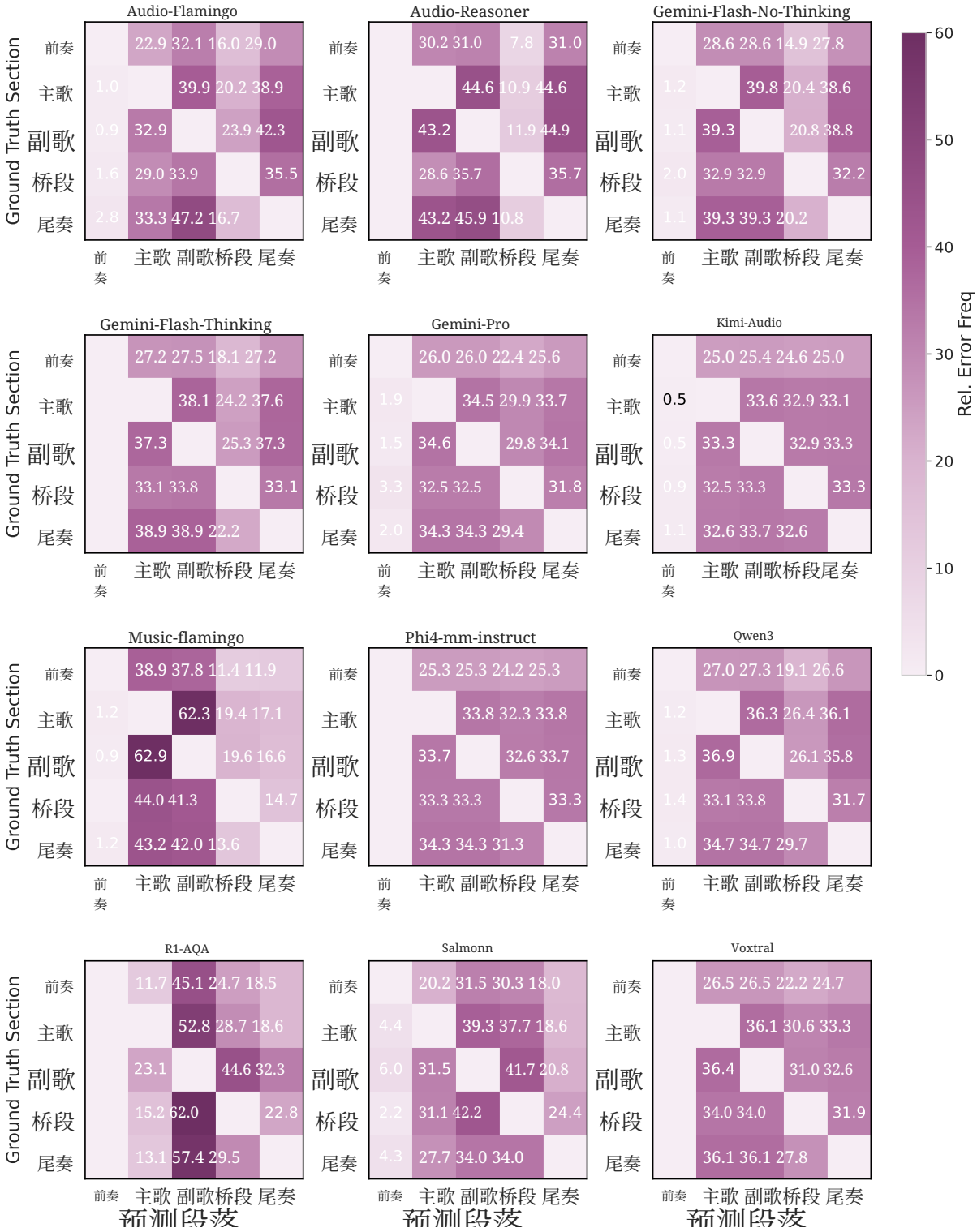


图16：各模型在结构分段中的段落预测错误，突出显示副歌与主歌乐段之间的频繁混淆。