arXiv:2601.07237v1 [eess.AS] 12 Jan 2026

ICASS P 2026 自动歌曲美学评估挑战赛 E

Guobin Ma[1] *Yuxuan Xia*[1] *Jixun Yao*[1] *Huixin Xue*[2] *Hexin Liu*[3] *Shuai Wang*[4] *Hao Liu*[2] *Lei Xie* [1*]

[1] 西北工业大学计算机科学学院音频、语音与语言处理组 (ASLP@NPU)，中国西安[2] 上海音乐学院，中国上海[3] 南洋理工大学计算与数据科学学院，新加坡[4] 南京大学，中国苏州

摘要

本文总结了 ICASSP 2026 自动歌曲美学评估（ASAE）挑战赛[1], , 该挑战赛聚焦于预测人工智能生成歌曲的主观美学评分。挑战赛设有两个赛道：赛道1旨在预测总体音乐性评分，赛道2则聚焦于预测五项细粒度美学评分。该挑战赛受到研究界的广泛关注，并吸引了来自学术界和产业界的大量投稿。表现最佳的系统显著超越了官方基线，展现出客观指标与人类审美偏好相契合方面的显著进展。挑战赛成果建立了一个标准化基准，并推动了面向现代音乐生成系统、以人为本的评估方法的发展。

索引词— 音乐评估，ASAE挑战赛，SongEval

1. 引言

随着生成式人工智能的快速发展，音乐生成已从仅能产出连贯的音频波形，迈向创作能够引发听众情感共鸣的艺术内容。商业音乐生成器（如 Suno [1], 、Udio [2], 和 Mureka [3]）通过生成高保真歌曲，缩小了人类与机器创造力之间的差距，改变了创作过程。然而，评估生成歌曲的主观美学表现，并捕捉情感表达、音乐性和听众愉悦度等方面，仍是一项重大挑战。诸如 Fr´echet 音频距离 [4]之类的客观评估指标能够捕捉分布距离，但与审美感知的相关性较弱 [5, 6]。Meta Audiobox Aesthetics 等无参考模型 [7] 能够泛化至不同音频领域，但更重视声学保真度，而非高层次的音乐美学。

为弥补这一空白，我们组织了 ICASSP 2026 ASAE挑战赛，旨在建立一个标准化基准，用于对生成歌曲进行无参考审美评估，并聚焦于以人类听众为中心的标准。挑战赛包含两个赛道：整体音乐性预测（赛道1），侧重于整体评估；以及细粒度维度预测（赛道2），将评估扩展到五个具体的审美维度。

*通讯作者。[1]https://aslp-lab.github.io/Automatic-Song-Aesthetics-Evaluation-Challenge/

2. 数据集

为支持本次挑战赛，我们采用 SongEval 基准 [8],, 使用其数据集进行训练，并将基于 UTMOS 的 [9] 模型作为官方基线[2]。此外，我们构建了专门的测试集，以严格评估模型性能；其中所有歌曲均由具备正规音乐背景的专家标注。

赛道1包含两个评测子集：常规集包含由5个已见系统（Suno [1], Udio [2],Mureka [3], DiffRhythm [10], YUE [11]）生成的歌曲，用于评估分布内性能；困难集则包含这5个已见系统以及6个未见系统（Seed-Music [12], ElevenLabs [13], ACE-step [14],JAM [15], SongGeneration [16], Songbloom [17]）生成的歌曲，用于评估鲁棒泛化。

赛道2采用专门构建并精心筛选的测试集，针对五个美学维度：整体连贯性、可记忆性、人声自然度、结构清晰度和整体音乐性。该数据集确保广泛覆盖各类美学特质，并包含存在孤立退化的样本（例如，可记忆性高但连贯性低），从而能够对维度特异性敏感度进行细粒度评估。

3. 挑战赛指标

我们提出了一种分层评估指标，用于评估话语级 （单首歌曲）和系统级 （每个模型的平均得分）表现。

3.1. 赛道1：整体音乐性预测

赛道1侧重于预测一个代表整体审美质量的单一标量得分。为了测试模型的泛化能力，我们采用了第2节中介绍的两个不同测试集：**测试集1**（常规）和**测试集2**（困难）。对于每个测试集，我们通过汇总四项指标计算综合得分 S_{set}：皮尔逊线性相关系数（LCC） [18], 、斯皮尔曼等级相关系数（SRCC） [19],、肯德尔秩相关系数（KATU） [20], 以及顶级准确率（TTA[3]）。请注意，TTA仅在话语级计算。综合得分定义如下：

$$S_{set} = \frac{1}{4}\left(\text{TTA}_u + \sum_{m\in\mathcal{M}} \frac{m_u + m_s}{2}\right) \tag{1}$$

其中， $\mathcal{M} = \{\text{LCC、SRCC、KATU}\}$以及下标 u 和 s 分别表示话语级和系统级。

[2]https://github.com/ASLP-lab/SongEval[3]TTA 衡量正确分类顶级歌曲的准确率。

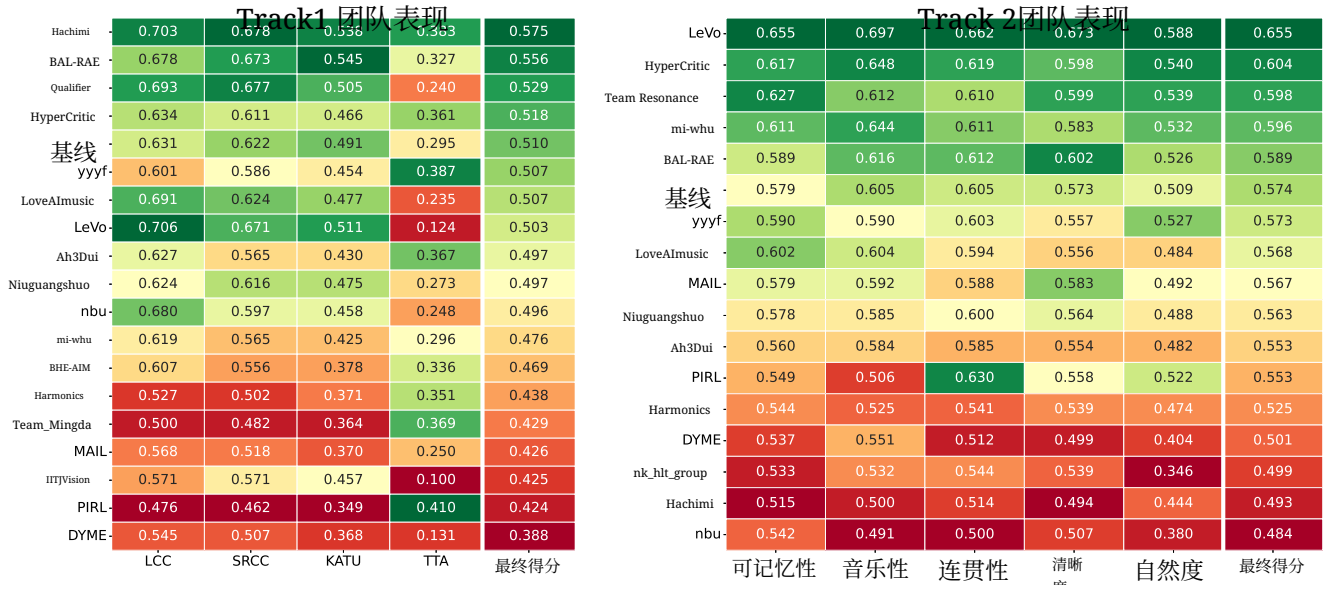

图 1。挑战结果概览。图中展示了参与团队在赛道1（整体音乐性）和赛道2（细粒度维度）中的表现分布，并与基线系统进行了比较。

为增强稳健性，赛道1的最终排名由优先考虑困难集的加权和决定：

$$\text{Score}_{\text{Track1}} = 0.2 \times S_{\text{set1}} + 0.8 \times S_{\text{set2}} \tag{2}$$

3.2. 赛道2：细粒度维度预测

赛道2要求参与者同时预测全部五个审美维度的得分。评估使用第2节所述的专用赛道2测试集进行。

对于每个维度 $d \in \{1, \ldots, 5\}$，我们采用与公式 (1) 相同的复合指标公式计算维度得分 S_d。赛道2的最终排名指标为所有审美维度的宏平均：

$$\text{Score}_{\text{Track2}} = \frac{1}{5} \sum_{d=1}^{5} S_d \tag{3}$$

该评测方案更青睐在所有审美方面均表现稳定的模型，而非偏重特定维度的模型。

4. 结果与结论

本次挑战赛受到社区广泛关注，注册参赛的团队超过70支。最终，我们分别收到赛道1和赛道2的18份和16份有效提交。图1总结了参赛团队在两个赛道中的表现。

对于赛道1，表现最佳的系统始终胜过官方基线，并且大多数系统都采用了相似策略：利用多种预训练音频表征，并对歌曲的长程结构进行显式建模。获胜团队***Hachimi***使用双阶段双向交叉注意力模块融合了 MuQ [21] 和 WavLM [22]嵌入，使音乐结构线索与声学细节能够相互交换信息，这对整体音乐性预测尤其有益。***BAL-RAE***通过多尺度特征提取和混合排序感知优化策略，令训练目标与竞赛的相关性和TTA指标保持一致。资格赛选手将大规模音乐预训练与基于度量的元学习框架相结合，有效缓解了标签有限情况下的过拟合，并提升了对困难划分的泛化能力。***HyperCritic*** 也表现出色，它将异构SSL骨干网络与结构感知分支（如 SongFormer[23]）相结合，进一步验证了结构建模对整体审美的重要性。

对于赛道2，表现突出的参赛作品侧重于维度特定建模，并力求在各项属性上取得均衡表现。获胜团队 ***LeVo*** 采用混合专家融合策略，对多种预训练骨干网络进行融合。该方法能够自适应地为互补线索赋权——从声学特征到生成式特征——从而在五个维度上取得最稳定的分数。包括***HyperCritic***和***TeamResonance***在内的其他优秀团队也进一步彰显了多样化方法的优势：前者通过统一Transformer [24],整合了多路特征流，后者则采用稳健回归集成来降低方差。***mi-whu*** 团队采用双编码器设计，在共享通用音乐因素的同时保留维度特定池化，取得了有竞争力的结果。结合数据增强和轻量级微调，该策略在不牺牲稳定性的前提下提升了细粒度敏感性。

总体而言，这些结果表明，在使客观指标与人类审美判断保持一致方面取得了明显进展。表现领先的系统凸显了多表征融合、结构感知建模和相关性对齐目标的有效性。然而，一个长期存在的瓶颈依然未能解决：尽管相关性总体呈现强劲趋势，各团队的 TTA 得分普遍不高。这种差异表明，模型虽然能够成功捕捉整体审美排名，却难以在高质量区间进行细粒度区分，常常低估顶级曲目。我们希望本次挑战赛能够推动音乐生成领域的进步，同时，配套的评估模型也能助力系统优化，并在强化学习框架中充当奖励函数。

5. 参考文献

[1] Suno，“Suno 官方网站”，*Suno*，2025年，https://suno.com。[2] Udio，“Udio 官方网站”，*Udio*，2025年，https://www.udio.com/home。[3] Mureka，“Mureka 官方网站”，*Mureka*，2025年，https: //www.mureka.ai。[4]Kevin Kilgour、Mauricio Zuluaga、Dominik Roblek 和 Matthew Sharifi，“Fr´echet 音频距离：一种用于评估音乐增强算法的无参考指标”，载于 *IN-TERSPEECH*。2019年，第2350–2354页，ISCA。[5] Ashvala Vinay 和 Alexander Lerch，“生成式音频系统及其指标的评估”，载于 *ISMIR*，2022年，第858–865页。[6] Modan Tailleur、Junwon Lee、Mathieu Lagrange、Keunwoo Choi、Laurie M. Heller、Keisuke Imoto 和 Yuki Okamoto，“Fr´echet 音频距离与环境音频的人类感知之间的相关性取决于嵌入”，载于 *EU- SIPCO*。2024年，第56–60页，IEEE。[7] Andros Tjandra、Yi-Chiao Wu、Baishan Guo、John Hoffman、Brian Ellis、Apoorv Vyas、Bowen Shi、Sanyuan Chen、Matt Le、Nick Zacharov、Carleigh Wood、Ann Lee 和 Wei-Ning Hsu，“Meta audiobox aesthetics：面向语音、音乐和声音的统一自动质量评估”，*CoRR*，第abs/2502.05139卷，2025年。[8] Jixun Yao、Guobin Ma、Huixin Xue、Huakang Chen、Chunbo Hao、Yuepeng Jiang、Haohe Liu、Ruibin Yuan、Jin Xu、Wei Xue 等，“Songeval：用于歌曲审美评估的基准数据集”，*abs/2505.10793*，2025年。[9] Takaaki Saeki、Detai Xin、Wataru Nakata、Tomoki Koriyama、Shinnosuke Takamichi 和 Hiroshi Saruwatari，“UTMOS：UTokyo-SaruLab 面向 VoiceMOS 挑战赛 2022 的系统”，载于 *IN- TERSPEECH*。2022年，第4521–4525页，ISCA。[10] Ziqian Ning、Huakang Chen、Yuepeng Jiang、Chunbo Hao、Guobin Ma、Shuai Wang、Jixun Yao 和 Lei Xie，“DiffRhythm：以潜扩散实现快速且极为简单的端到端全长歌曲生成”，*abs/250301183*，2025年。。[11] Ruibin Yuan、Hanfeng Lin、Shuyue Guo、Ge Zhang、Jiahao Pan、Yongyi Zang、Haohe Liu、Yiming Liang、Wenye Ma、Xingjian Du、Xinrun Du、Zhen Ye、Tianyu Zheng、Yinghao Ma、Minghao Liu、Zeyue Tian、Ziya Zhou、Liumeng Xue、Xingwei Qu、Yizhi Li、Shangda Wu、Tianhao Shen、Ziyang Ma、Jun Zhan、Chunhui Wang、Yatian Wang、Xiaowei Chi、Xinyue Zhang、Zhenzhu Yang、Xiangzhou Wang、Shansong Liu、Lingrui Mei、Peng Li、Junjie Wang、Jianwei Yu、Guo-jian Pang、Xu Li、Zihao Wang、Xiaohuan Zhou、Lijun Yu、Emmanouil Benetos、Yong Chen、Chenghua Lin、Xie Chen、Gus Xia、Zhaoxiang Zhang、Chao Zhang、Wenhu Chen、Xinyu Zhou、Xipeng Qiu、Roger B. Dannenberg、Zheng-Jia Liu、Jian Yang、Wenhao Huang、Wei Xue、Xu Tan 和 Yike Guo，“YUE：扩展开放基础模型以生成长篇音乐”，*CoRR*，第abs/2503.08638卷，2025年。

[12] Seed-Music，“Seed-music 官方网站”，*Seed-Music*，2025年，https://www.doubao.com/chat。

[13] ElevenLabs，“Elevenlabs 官方网站”，*ElevenLabs*，2025年，https://elevenlabs.io/app/music。

[14] Junmin Gong、Sean Zhao、Sen Wang、Shengyuan Xu 和 Joe Guo，“Ace-step：迈向音乐生成基础模型”，*CoRR*，第 abs/2506.00045 卷，2025年。[15]Renhang Liu、Chia-Yu Hung、Navonil Majumder、Taylor Gautreaux、Amir Ali Bagherzadeh、Chuan Li、Dorien Herremans 和 Soujanya Poria，“JAM：一款具备细粒度可控性和美学对齐能力的轻量级基于流的歌曲生成器”，*CoRR*，第 abs/2507.20880 卷，2025年。[16] Shun Lei、Yaoxun Xu、Zhiwei Lin、Huaicheng Zhang、Wei Tan、Hangting Chen、Jianwei Yu、Yixuan Zhang、Chenyu Yang、Haina Zhu、Shuai Wang、Zhiyong Wu 和 Dong Yu，“Levo：通过多偏好对齐实现高质量歌曲生成”，*CoRR*，第 abs/2506.07520 卷，2025年。[17] Chenyu Yang、Shuai Wang、Hangting Chen、Wei Tan、Jianwei Yu 和 Haizhou Li，“Songbloom：通过交错自回归草图生成和扩散细化实现连贯歌曲生成”，*CoRR*，第 abs/2506.07634 卷，2025年。[18] Philip Sedgwick，“皮尔逊相关系数”，*Bmj*，第 345 卷，2012年。[19] Philip Sedgwick，“斯皮尔曼等级相关系数”，*Bmj*，第 349 卷，2014年。[20] A Ian McLeod，“肯德尔秩相关与曼-肯德尔趋势检验”，*R* 包 *Kendall*，第 602 卷，第 1–10 页，2005年。[21] Haina Zhu、Yizhi Zhou、Hangting Chen、Jianwei Yu、Ziyang Ma、Rongzhi Gu、Yi Luo、Wei Tan 和 Xie Chen，“Muq：采用梅尔残差向量量化的自监督音乐表征学习”，*CoRR*，第 abs/2501.01108 卷，2025年。[22] Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao、Jian Wu、Long Zhou、Shuo Ren、Yanmin Qian、Yao Qian、Jian Wu、Michael Zeng、Xiangzhan Yu 和 Furu Wei，“Wavlm：面向全栈语音处理的大规模自监督预训练”，*IEEEJ.Sel.Top. SignalPro- cess.*，第 16 卷，第 6 期，第 1505–1518 页，2022年。[23] Chunbo Hao、Ruibin Yuan、Jixun Yao、Qixin Deng、Xinyi Bai、Wei Xue 和 Lei Xie，“Songformer：通过异构监督扩展音乐结构分析”，*arXiv preprint arXiv:2510.02797*，2025年。[24] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin，“注意力就是一切”，发表于*NIPS*，2017年，第 5998–6008 页。