

Academic Text-to-Music Grand Challenge: Datasets, Baselines, and Evaluation Methods

Fang-Chih Hsieh^{1*}, Wei-Jaw Lee¹, Chun-Ping Wang¹, Hung-yi Lee¹, Hao-Wen Dong², and Yi-Hsuan Yang¹

¹Artificial Intelligence Center of Research Excellence, National Taiwan University, Taipei, Taiwan

²Department of Performing Arts Technology, University of Michigan, Ann Arbor, MI, United States

Abstract—This paper presents an overview and the technical framework of the ICME 2026 Grand Challenge on Academic Text-to-Music Generation (ATTM). Despite the rapid progress in text-to-music generation (TTM) systems, the field is currently dominated by models trained on massive proprietary datasets with industrial-scale computational resources, creating a significant barrier for academic research. To address this, the ATTM Challenge establishes a fair-play benchmark that requires participants to train generative models strictly from scratch using a standardized, CC-licensed subset of the MTG-Jamendo dataset containing only instrumental music. The challenge is divided into two tracks: the Efficiency Track (limited to 500M parameters) and the Performance Track (no parameter limit). Submissions are evaluated through a multi-stage process involving objective metrics, including Fréchet Audio Distance, CLAP score, and a novel Concept Coverage Score (CCS), followed by a subjective listening test. By providing open-source baselines, preprocessing pipelines, reference captions, and public evaluation code for computing FAD and CLAP, this challenge aims to facilitate and promote TTM research in academic contexts.

Index Terms—Text-to-music generation, generative AI, openness, reproducibility, affordability, large audio-language models

I. INTRODUCTION

The landscape of generative audio has shifted toward text-to-music (TTM) systems, which use architectures such as latent diffusion and Transformers to synthesize high-fidelity music from natural language. These models offer transformative potential for creators and educators, yet progress is increasingly stalled by a compute and data wall. State-of-the-art (SOTA) systems typically rely on massive, proprietary datasets and industrial-scale hardware, creating a significant barrier for the broader academic community. Researchers in academic labs often find themselves limited to fine-tuning existing models or performing small-scale experiments. This lack of access to standardized training conditions hampers the transparency, reproducibility, and fundamental architectural innovation necessary for the field to mature. Lee et al. [1] investigated this question by training a few TTM models from scratch under constrained data and computational resource. It motivates us to more comprehensively explore model capabilities in academic TTM settings.

In this paper, we present the ICME 2026 Grand Challenge on Academic Text-to-Music (ATTM) Generation.¹ The ATTM

challenge is designed as a “fair-play” benchmark to foster innovation under transparent and reproducible conditions. The core principle of this challenge is the requirement that all generative models must be trained from scratch using a standardized, CC-licensed dataset of 3,777 hours derived from the MTG-Jamendo corpus [2]. By restricting the data source and prohibiting the use of synthetic data generated by commercial engines, the challenge shifts the focus away from data scale and toward algorithmic efficiency, musical intelligence, and effective representation learning.

The challenge is structured into two distinct tracks to accommodate different resource levels. The Efficiency Track imposes a strict limit of 500M parameters on the core generative model, encouraging the development of lightweight yet powerful architectures. The Performance Track offers no parameter limits, challenging participants to push the upper bounds of musical quality within the provided data constraints. In both tracks, the task is restricted to generating 10-second instrumental music clips, which standardizes the evaluation target and keeps the benchmark practical for academic training and testing. To ensure a holistic assessment, we employ a multi-stage evaluation pipeline. Submissions are first screened using objective metrics, including Fréchet Audio Distance (FAD) [3], CLAP scores [4], [5], and a novel Concept Coverage Score (CCS) that utilizes large audio language models (LALMs) to verify the presence of specific musical attributes. Then, top-performing systems undergo a formal mean opinion score (MOS) study conducted by listeners to evaluate musicality and prompt adherence, and to determine the final ranking.

The contributions of the challenge are four-fold:

- 1) We establish a standardized fair-play benchmarking framework for TTM, providing a curated 3,777-hour dataset of instrumental music alongside automated vocal separation and captioning pipelines to ensure research transparency and reproducibility.
- 2) We introduce Concept Coverage Score (CCS), a novel evaluation methodology that uses LALMs to provide a fine-grained, interpretable assessment of semantic alignment between musical concepts and generated audio.
- 3) We curate in-distribution (ID) and out-of-distribution (OOD) prompt sets featuring seen and unseen tag combinations for a structured evaluation benchmark, enabling systematic analysis of model performance on compositional generalizability.

*Corresponding author: andrew891221@gmail.com

¹<https://ntu-musicailab.github.io/ICME26-ATTM-Grand-Challenge/>

- 4) We provide the open-source FluxAudio [6] baseline system and a suite of training scripts, fostering accessibility and lowering the entry barrier for academic teams.

II. TASK DEFINITION AND TRACK SPECIFICATIONS

A. Key Principles and Constraints

All participants must adhere to the following principles:

- **Training from scratch:** The core generative model responsible for the TTM mapping must be trained entirely from scratch. The use of pre-trained weights for the main generation architecture is strictly prohibited.
- **Exclusive data usage:** Participants must exclusively use the provided subset of the MTG-Jamendo dataset [2] for all training and data augmentation.
- **Prohibition of data laundering:** The use of external music datasets or synthetic audio generated by proprietary commercial models (e.g., Suno, Udio) is forbidden. Such practices are classified as data laundering and will result in immediate disqualification.
- **Auxiliary component policy:** While the core model must be trained from scratch, participants can use publicly available checkpoints for auxiliary components including:
 - Audio tokenizers, autoencoders (e.g., EnCodec [7]).
 - LALMs for automated captioning/tagging.
 - Vocoders or audio enhancement modules for post-processing.

However, proprietary or non-reproducible auxiliary models are prohibited.

- **No human-in-the-loop:** All generations must be fully autonomous, without any form of manual editing, human annotation, or cherry-picking of submitted samples.

B. Track Specifications

The challenge is divided into two tracks to accommodate varying research focuses and hardware availability.

1) *Efficiency Track:* This track is designed to encourage the development of compact and computationally efficient architectures. The core generative model is restricted to a maximum of 500M parameters. This limit excludes auxiliary components such as text encoders, audio decoders, or vocoders. This track is particularly suited for student teams and labs focusing on edge-AI or resource-constrained optimization.

2) *Performance Track:* This track has no parameter limits for the generative model. It challenges participants to push the boundaries of musical quality and semantic alignment using the provided academic dataset, allowing for the exploration of large-scale architectures or complex ensemble methods.

C. Definition of Core Generative Model

For the purpose of parameter counting and training restrictions, the “core generative model” refers to the architecture responsible for the mapping from the conditioned text representation (or latent tokens) to the musical representation. It can be a latent diffusion model, a Transformer-based decoder, or a masked generative model. Encoders used solely for feature extraction and decoders used for final waveform reconstruction are considered as auxiliary.

III. DATASET AND PREPROCESSING

The challenge focuses on instrumental music generation, excluding singing voices. We present below the curation of the dataset, the vocal separation pipeline, and the dual-model captioning strategy used to provide diverse semantic labels.

A. Source Dataset

The challenge utilizes the MTG-Jamendo dataset [2], which is an open-source music dataset built from tracks published on Jamendo under Creative Commons (CC) licenses, making it suitable for transparent and reproducible academic research. Specifically, we use the `raw_30s` subset, which contains 55,701 tracks with duration longer than 30 seconds. An additional advantage of this dataset is its per-track high-quality “tags” annotated by human experts. These tags are organized into three categories—genre, instrument, and mood/theme—comprising 226, 145, and 224 unique tags, respectively. As we focus on instrumental music generation, we require participants to transform this source data into a vocal-free corpus using a standardized preprocessing pipeline.

B. Vocal Separation and Preprocessed Data Size

We release the complete preprocessing codebase and require participants to perform vocal separation on their machine.

The preprocessing uses the Mel-Band Roformer model [8], a SOTA architecture for music source separation. Participants utilize the `melband-roformer-kim-vocals` checkpoint to isolate and remove vocal stems. To accommodate participants with limited time or computational resources, we provide an optional script to crop audio files to a shorter duration (30 seconds) prior to separation, which significantly reduces the total processing time. As a result, participants will obtain either the *full* vocal-removed Jamendo dataset, spanning around 3,777 hours and occupying roughly 240 GB, or a *partial* vocal-removed subset containing only the first 30 seconds of each track, spanning around 464 hours and substantially smaller at around 25 GB.

C. Audio Captioning

We offer two distinct reference caption sets generated by two different LALMs to provide semantic diversity. Participants may use these captions directly as a baseline training set, or utilize the provided captioning codebase to explore data augmentation strategies, or other captioning models.

1) *Pipeline A (“Qwen2-Audio direct captioning”):* uses Qwen2-Audio-7B-Instruct [9] to generate descriptive English captions focusing on genre, instrumentation, and mood. This pipeline produces holistic descriptions of the music’s atmosphere, disregarding any vocal elements.

2) *Pipeline B (“Music Flamingo with refinement”):* employs a two-stage approach to leverage the technical descriptive capabilities of the Music Flamingo model [10]. Since the raw output from Music Flamingo tends to be overly verbose, we employ a second model, Qwen3-4B-Instruct, to refine and rephrase the initial descriptions. With this refinement,

TABLE I
PROMPTS FOR REFERENCE CAPTION GENERATION

Model / Role	System Prompt / Instruction
Qwen2-Audio (Direct)	“Provide a detailed English caption for this music piece, focusing on aspects such as genre, instruments used, overall mood, and production style. Do not mention anything about singing, lyrics, or vocal.”
Music Flamingo (Initial)	“Describe this track in less than 3 sentences, focusing on aspects such as genre, instruments used, overall mood, tempo and key.”
Qwen3-4B (Refining)	“Rephrase the following music caption to be shorter and more concise while keeping all essential information about genre, instruments, mood, and production style. Remove any mention of vocals, lyrics, singers, voice types, or singing techniques. Also remove any mention of duration or length of the track. The rephrased caption must describe only instrumental content and production characteristics, and should be fluent and natural-sounding. Output only the rephrased caption, with no extra text, notes, or explanations.”

the captions remain concise, natural-sounding, and strictly instrumental in focus.

The specific prompts and instructions for both pipelines are summarized in Table I. The inclusion of both pipelines serves to encourage participants to investigate how caption diversity and LALM-based augmentation impact final performance.

D. Resource Accessibility

For transparency and lower entry barrier, the repositories for both the vocal separation pipeline² and the music captioning pipelines³ are publicly available on GitHub.

IV. TOPLINE AND BASELINE MODELS

Besides the submissions from the participants, we evaluate two categories of models: established pre-trained systems serving as the topline, and our official challenge baseline.

1) *Topline Models*: To provide a performance ceiling, we evaluate several SOTA TTM systems, including **Stable Audio Open** (SAO) [11], the **MusicGen** family (small, medium, and large) [12], and the **MeanAudio-Full** series (small and large) [13]. We use the official checkpoints for these models, which have been trained on massive proprietary or large-scale open datasets and serve as a reference for the current upper bound of unconstrained TTM performance.

2) *Official Baseline*: Our official baseline is **FluxAudio-S** (120M parameters), featuring a Flux-style Transformer architecture trained using the conditional flow matching (CFM) objective [6].⁴ Instead of using pretrained checkpoints, we train this baseline from scratch using the same vocal-removed MTG-Jamendo dataset provided to the participants and the reference captions generated by Qwen2-Audio-7B-Instruct [9]. It utilizes a pre-trained EnCodec decoder [7] for audio reconstruction and a T5-based text encoder [14] for prompt

²<https://github.com/ntu-musicalab/ICME26-ATTM-GC-Preprocessing>

³<https://github.com/ntu-musicalab/ICME26-ATTM-GC-ALM-captioning>

⁴According to the original paper [13], MeanAudio is distilled from FluxAudio for single-step generation through consistency-based objectives. We use the FluxAudio architecture for our baseline to prioritize generative quality.

conditioning. In our implementation, the baseline model was trained on a single NVIDIA RTX A6000 with 48GB VRAM. The training proceeded for 200,000 steps with a batch size of 128. The total training time was about 2 days and 4 hours. We shared the codebase of this baseline model at the outset of the competition, again to lower the entry barrier.⁵

V. EVALUATION METHODOLOGY

The evaluation is divided into two phases. Phase 1 utilizes objective metrics to establish a quantitative scorecard for all submissions. Phase 2 uses a subjective MOS study to provide a qualitative assessment of the top-performing systems.

A. Evaluation Data Curation

To ensure a rigorous evaluation, we curate a set of 100 test prompts through a multi-stage pipeline involving tag filtering, stratified sampling, and LLM synthesis. This process ensures that the test prompts are musically plausible, semantically diverse, and distributionally consistent with the training data.

1) *Tag pool filtering*: We utilize the original tag metadata from the MTG-Jamendo `raw_30s_clean-tags.csv` as the initial tag pool. To ensure that the evaluation focuses on well-defined and verifiable musical concepts, we filtered the tags based on the following four criteria:

- 1) Popularity: Each tag must appear in at least 100 tracks within the original dataset.
- 2) LALM verifiability: Using the Qwen3-Omni model [15] (the judge for our CCS metric, see Section V-B), each tag must achieve a recall rate ≥ 0.85 based on the MTG-Jamendo ground truth.
- 3) Reference consistency: Each tag must appear at least 10 times across the two reference caption sets we provide.
- 4) Instrumental constraint: All vocal-related tags such as “choir” and “vocals” are manually excluded.

This filtering leads to a final taxonomy of 143 valid tags, comprising 55 genres, 25 instruments, and 63 moods/themes.

2) *Stratified triplet sampling*: From the filtered taxonomy, we sampled 100 unique triplets of tags, each consisting of exactly one genre, one instrument, and one mood/theme tag. To assess the models’ generalization capabilities while maintaining fairness, we categorized these triplets based on their co-occurrence patterns in the original dataset:

- **In-Distribution (ID)**: 80 triplets were sampled where all tag pairs in a triplet (i.e., “genre & instrument”, “genre & mood/theme”, and “instrument & mood/theme”) have co-occurred at least once in the training data, not necessarily in the same songs.
- **Out-of-Distribution (OOD)**: 20 triplets were sampled where at least one tag pair in a triplet has never co-occurred (e.g., “heavy metal” & “calm”).

While OOD samples are included to explore model behavior in edge cases for future work, only scores achieved on the 80 ID prompts contribute to the final official leaderboard ranking. The reason is that our objective evaluation metrics have not

⁵<https://github.com/ntu-musicalab/ICME26-ATTM-GC-FluxAudio>

yet been verified on the OOD samples, and therefore may not produce sufficiently reliable scores for official ranking.

3) *Prompt synthesis via LLM*: The sampled triplets were transformed into fluent, descriptive English captions using Qwen3-4B-Instruct. To prevent distributional shift between the test prompts and the provided training captions, we employ a 10-shot in-context learning (ICL) approach, providing the model with random examples/demonstrations from the reference caption sets (5 shots from captioning pipeline A and 5 shots from pipeline B) to match the tone and phrasing of the employed music captioning pipelines.

Notably, we apply two synthesis strategies for the triplets:

- 1) **Strict following** (40 ID and 20 OOD): The model is instructed to describe the music using exclusively the provided tags, not adding other musical elements.
- 2) **Improvisation** (40 ID): To improve musical coherence and simulate realistic descriptions, the model is prompted to improvise by adding one to three additional, musically plausible instruments to the caption. This improvisation is only used to enrich the prompt text: the enriched prompt is used for computing the CLAP score, but the added instruments are not counted as target concepts in the CCS evaluation.

The resulting 100 prompts serve as the final evaluation set, providing a balanced mix of strict tag adherence and naturally enriched musical descriptions.

B. Objective Evaluation (Phase 1)

In the first phase, submissions are ranked using a composite objective score. Each system is evaluated based on 10-second generated audio clips, matching the challenge task definition. While standard metrics such as FAD and CLAP scores are employed, we emphasize the use of high-correlation embedders and introduce a novel granular metric named CCS for semantic verification. To facilitate transparent and reproducible benchmarking, we also release the official evaluation code for computing FAD and CLAP to the public.⁶

1) **Fréchet Audio Distance**: FAD evaluates the distributional similarity between generated audio and a reference set of real music. We adopt the LAION-CLAP-Music model `music_audioset_epoch_15_esc_90.14` as the feature extractor, as it provides FAD scores most aligned with human preference according to [16], [17]. For the reference set, we use a hidden instrumental subset from MTG-Jamendo consisting of 1,000 randomly sampled tracks.

2) **CLAP Score**: To measure the global semantic alignment between the input prompt y and the generated audio \mathbf{x} , we compute the cosine similarity in the joint embedding space of the CLAP model. Following [17], we employ the same `music_audioset_epoch_15_esc_90.14` checkpoint, which exhibits the highest correlation with human ratings regarding text-to-audio relevance in the music domain.

3) **Concept Coverage Score (CCS)**: While the CLAP score provides a global measure of semantic similarity, it lacks interpretability regarding specific musical attributes. To address this, we propose CCS, a novel metric that utilizes the SOTA LALM, Qwen3-Omni [15], as a zero-shot music judge to verify the presence of individual musical tags. Specifically, given a test prompt synthesized from a triplet of tags $T = \{t_g, t_i, t_m\}$ representing genre, instrument, and mood/theme respectively, CCS assesses how many of these target concepts can be detected from the generated audio \mathbf{x} .

a) *LALM-based verification*: For each concept $t \in T$, LALM is provided with the audio \mathbf{x} and a category-specific prompt. These prompts instruct the model to act as a specialized classifier (e.g., a music genre classifier or an audio event detector) and determine if any trace of the concept t is present. To ensure robustness and bypass potential linguistic biases in text generation, we do not rely on the model’s textual output. Instead, we extract the log-probabilities (logits) for the tokens “Yes” and “No.” The detection function $D(\mathbf{x}, t)$ is defined as:

$$D(\mathbf{x}, t) = \begin{cases} 1 & \text{if } \text{logit}(\text{“Yes”}) > \text{logit}(\text{“No”}), \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

b) *Scoring and aggregation*: We report a single aggregate CCS that measures the fraction of target concepts detected across all evaluation samples. Specifically, for a set of N evaluation samples $\{\mathbf{x}_i, T_i\}_{i=1}^N$, the score is calculated as:

$$\text{CCS} = \frac{1}{3N} \sum_{i=1}^N \sum_{t \in T_i} D(\mathbf{x}_i, t). \quad (2)$$

C. Final Ranking of the Objective Result via Borda Count

We employ the Borda count method to convert the scores in the three objective metrics into a single leaderboard. Specifically, for a ranking pool with M submissions, we rank all submissions plus the FluxAudio baseline (i.e., there are in total $M + 1$ candidates) for each objective metric independently, in ascending order for FAD, and in descending order for CLAP and CCS. Let $r_m(s) \in \{1, \dots, M + 1\}$ denote the rank of submission s under metric $m \in \{\text{FAD}, \text{CLAP}, \text{CCS}\}$, where rank 1 is the best. The Borda score assigned to submission s for metric m is then $B_m(s) = M + 1 - r_m(s)$. Thus, the best submission for a metric receives M points (i.e., the higher the better), the second-best receives $M - 1$ points, etc, until the lowest-ranked submission receives 0 points.

The overall objective score of submission s is the sum of its Borda scores across the three metrics:

$$B_{\text{total}}(s) = B_{\text{FAD}}(s) + B_{\text{CLAP}}(s) + B_{\text{CCS}}(s). \quad (3)$$

Submissions are then sorted by $B_{\text{total}}(s)$ to obtain the objective ranking. The toplines, baseline and submissions are all evaluated on the 80 ID test prompts curated in Section V-A.

Since each team may submit up to two results per track, we use a two-stage ranking procedure. In the first stage, we run the Borda count separately within the Efficiency Track and the Performance Track. If a team has two submissions in the same

⁶<https://github.com/ntu-musicailab/ICME26-ATTM-GC-Evaluation>

TABLE II
COMPREHENSIVE BENCHMARKING SUMMARY FOR THE OBJECTIVE EVALUATION USING THE FINAL TEST PROMPTS.

Model	Params	Train Data (hours)	Arch.	Official Caption	MTG Tags	Post-train	Infer. Opt.	GPU-H (hours)	FAD↓	CLAP↑	CCS↑	Rank
Submission e00 [18]	120M	0.46K	D/F, T	✓	×	×	×	384	0.556	<u>0.310</u>	0.796	6
Submission e01 [19]	189M	3.7K	D/F, SSM	✓	×	×	×	81	0.577	0.338	<u>0.863</u>	2
Submission e02 [20]	120M	3.7K	D/F, T	✓	×	✓	×	40	0.498	0.270	0.763	8
Submission e03	340M	0.46K	D/F	✓	✓	×	✓	60	0.518	0.251	0.763	12
Submission e04	70M	0.46K	D/F, T	✓	×	×	×	8	0.574	0.195	0.833	9
Submission e05 [21]	499M	0.46K	D/F, T	✓	×	✓	✓	20	0.487	0.305	0.800	2
Submission e06	488M	0.46K	T, SSM	✓	×	×	✓	12	0.667	0.268	0.808	9
Submission e07 [22]	402M	3.7K	D/F	×	×	✓	×	300	0.417	0.261	0.867	1
Submission e08 [23]	450M	3.7K	D/F, T	✓	×	×	×	60	0.495	0.295	0.804	2
Submission e09 [24]	480M	3.7K	D/F, T	✓	×	×	×	72	0.646	0.263	0.767	12
Submission e10 [25]	315M	0.46K	T	✓	×	×	×	152	<u>0.482</u>	0.163	0.738	11
Submission e11	480M	0.46K	D/F	✓	×	✓	×	72	0.892	0.097	0.675	16
Submission p00 [18]	502M	0.46K	D/F, T	✓	×	×	×	384	0.557	0.311	<u>0.796</u>	6
Submission p05 [21]	2.4B	0.46K	D/F, T	✓	×	✓	✓	60	<u>0.514</u>	<u>0.306</u>	0.800	5
Submission p09 [24]	480M	3.7K	D/F, T	✓	×	×	×	72	0.646	0.260	0.767	15
Submission p10 [25]	1.5B	0.46K	T	✓	×	×	×	416	0.500	0.171	0.721	14
FluxAudio-S (Baseline)	120M	3.7K	D/F, T	✓	×	×	×	52	0.757	0.088	0.592	17
Stable Audio Open [11]	1.1B	7.3K	D/F	–	–	–	–	22,000	0.574	0.321	0.800	–
MusicGen-small [12]	300M	20K	T	–	–	–	–	N/A	0.574	0.370	0.875	–
MusicGen-medium [12]	1.5B	20K	T	–	–	–	–	N/A	0.548	0.353	0.892	–
MusicGen-large [12]	3.3B	20K	T	–	–	–	–	N/A	0.553	0.379	0.888	–
MeanAudio-S-Full [13]	120M	10K	D/F, T	–	–	–	–	280	0.649	0.210	0.808	–
MeanAudio-L-Full [13]	480M	10K	D/F, T	–	–	–	–	N/A	0.660	0.202	0.783	–

Note: Submission letters indicate the challenge track: e denotes the Efficiency Track and p denotes the Performance Track. The “Train Data (h)” column reports the amount of training audio in thousands of hours; 3.7K indicates use of the full MTG-Jamendo dataset (about 3,777 hours), whereas 0.46K indicates use of the 30-second subset (approximately 464 hours). “Arch.” stands for architecture; “D/F” denotes diffusion/flow-matching, “T” denotes Transformer, and “SSM” denotes state space models. “Official Caption” indicates whether a system uses the two provided official caption sets; if not, the team curated an alternative caption set, for example through data augmentation. “MTG Tags” indicates whether the original MTG-Jamendo tags are used during training. “Post-train” denotes whether post-training techniques are used. “Infer. Opt.” denotes whether inference-time optimization techniques are used. “GPU-H” denotes the self-reported approximate total GPU-hours spent on training, disregarding the actual type of GPU used (so the values are not directly comparable and are only for reference). We highlight the best result in each track, and underscore the second best. “Rank” indicates the ranking of the objective result.

track, we retain only the higher-ranked one as that team’s final entry for that track. After this filtering step, each team now has at most one submission in each track.

In the second stage, we merge the remaining submissions from both tracks into a joint submission pool and run the Borda count again. The resulting scores are used to produce the official objective ranking and to determine which teams advance as finalists to the subjective evaluation phase. We select the top 6 submissions in this ranking as the finalists, requiring that every finalist beat the baseline.

This aggregation scheme has two advantages. First, it balances the three objective criteria without requiring manual tuning of metric weights. Second, it discourages over-optimization toward any single metric, since a strong final ranking requires good performance across fidelity, global text-audio alignment, and concept-level semantic coverage.

D. Objective Benchmarking Results

The challenge was officially launched on February 10, 2026, with the final test prompts released on April 20, 2026, and the final audio submission deadline on April 23, 2026. In total, 18 teams from around the world registered, and we finally received submissions from 12 teams for the Efficiency Track, among which 4 teams also submitted to the Performance

Track. Since some teams submitted two systems within a track, we retained only the higher-ranking one using the Borda count procedure described in Section V-C. Table II reports the final submissions, our baseline, and the topline, and summarizes their model size, training data, core-model architecture, design choices, and objective results.

All submissions to either track beat the FluxAudio baseline. Following the objective ranking, we select four teams from the Efficiency Track (e01, e05, e07, e08) and two teams from the Performance Track (p00, p05) as the finalists.

E. Subjective Evaluation (Phase 2)

For the listening test, we compare the 6 finalists chosen from the two tracks, along with a topline model. Specifically, we choose MusicGen-small [12], for it is the only MusicGen series with fewer than 500M parameters, making it compatible with our parameter limit. We note that this official MusicGen-small checkpoint was trained on “20K hours of licensed music” [12], which is at least five times larger than our MTG-Jamendo dataset. We view MusicGen-small as a high anchor, not requiring any finalist to outperform it.

Therefore, the listening test compares in total 7 model variants, including the official MusicGen-small and the 6 finalists. Each questionnaire contains 5 prompts randomly

sampled from the 100 test prompts (1 OOD and 4 ID, two of which involve improvisation), along with the 10-second clips generated by the 7 variants, yielding 35 evaluation items per questionnaire. For broader prompt coverage while keeping the listening load manageable, we form 5 independent questionnaires using different prompt subsets, so that each variant is evaluated on 25 samples in total across the full study. We distribute the questionnaires through mailing lists and ask volunteer evaluators to answer one of them, preserving matched comparisons among variants within each prompt while distributing listener effort across the questionnaires. The evaluators will rate the generated samples on a 5-point Likert scale (the higher is better) across the following 4 criteria:

- **Audio Fidelity:** Evaluates the technical production quality, signal clarity, and the absence of perceptible digital artifacts or distortion.
- **Prompt Adherence:** Assesses how accurately the generated musical elements, such as instrumentation and genre, correspond to the attributes requested in the text prompt.
- **Musicality:** Measures intrinsic musical quality, including rhythmic stability, harmonic coherence, structural development, and general listenability.
- **Overall:** Captures the holistic impression of the generated music by integrating production quality, prompt relevance, and musical appeal into a single overall judgment.

F. Subjective Evaluation Results

We conducted the subjective listening test and received 35 participant responses in total. Among these responses, 25 were provided by expert listeners. Expert listeners are defined as participants who have more than three years of musical background, currently work in a music-related profession, or rate their music appreciation level above 3. Table III reports the final subjective evaluation results based on the ‘‘Overall’’ MOS, along with (\pm) the standard deviations. We report two sets of MOS scores: MOS_all, which includes responses from all participants, and MOS_expert, which includes only responses from expert listeners.

VI. CONCLUSION

The ATTM Grand Challenge establishes a rigorous foundation for academic research in text-to-music generation. By standardizing the dataset, providing a transparent evaluation protocol through the novel CCS objective evaluation metric, and offering a reproducible FluxAudio baseline, we aim to encourage algorithmic breakthroughs that are independent of massive industrial-scale computational resources. We invite the community to participate in this effort to build more intelligent, efficient, and interpretable music generation systems.

ACKNOWLEDGMENT

This work is supported by grants from the Ministry of Education (MOE) of Taiwan (for Taiwan Centers of Excellence), the National Science and Technology Council of Taiwan (NSTC 114-2628-E-002-013-MY3), and Google Asia Pacific. We are also grateful for the support of Moises AI in providing cash awards for the challenge winners.

TABLE III
SUBJECTIVE EVALUATION RESULTS BASED ON OVERALL MOS SCORES

Model	MOS_all	MOS_expert	Award
Submission e01 [19]	3.225 \pm 1.093	3.177 \pm 1.136	Efficiency 2nd
Submission e05 [21]	2.969 \pm 1.194	2.929 \pm 1.230	–
Submission e07 [22]	3.250 \pm 1.234	3.186 \pm 1.286	Efficiency 1st
Submission e08 [23]	3.119 \pm 1.084	3.044 \pm 1.113	Efficiency 3rd
Submission p00 [18]	2.006 \pm 1.031	2.044 \pm 1.089	–
Submission p05 [21]	3.344 \pm 1.116	3.327 \pm 1.137	Performance 1st
MusicGen-small [12]	3.538 \pm 1.009	3.425 \pm 0.998	–

REFERENCES

- [1] Wei-Jaw Lee et al., ‘‘Training-efficient text-to-music generation with state-space modeling,’’ *arXiv preprint arXiv:2601.14786*, 2026.
- [2] Dmitry Bogdanov et al., ‘‘The MTG-Jamendo dataset for automatic music tagging,’’ in *ICML*, 2019.
- [3] Kevin Kilgour et al., ‘‘Fréchet Audio Distance: A metric for evaluating music enhancement algorithms,’’ *arXiv preprint 1812.08466*, 2019.
- [4] Benjamin Elizalde et al., ‘‘CLAP: Learning audio concepts from natural language supervision,’’ in *ICASSP*, 2023.
- [5] Yusong Wu et al., ‘‘Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,’’ *arXiv preprint arXiv:2211.06687*, 2024.
- [6] Zhengcong Fei et al., ‘‘FLUX that plays music,’’ *arXiv preprint arXiv:2409.00587*, 2024.
- [7] Alexandre Défossez et al., ‘‘High fidelity neural audio compression,’’ *arXiv preprint arXiv:2210.13438*, 2022.
- [8] Ju-Chiang Wang et al., ‘‘Mel-Band Reformer for music source separation,’’ *arXiv preprint arXiv:2310.01809*, 2023.
- [9] Yunfei Chu et al., ‘‘Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models,’’ *arXiv preprint arXiv:2311.07919*, 2023.
- [10] Sreyan Ghosh et al., ‘‘Music Flamingo: Scaling music understanding in audio language models,’’ *arXiv preprint arXiv:2511.10289*, 2025.
- [11] Zach Evans et al., ‘‘Stable Audio Open,’’ in *ICASSP*, 2025.
- [12] Jade Copet et al., ‘‘Simple and controllable music generation,’’ *NeurIPS*, 2023.
- [13] Xiquan Li et al., ‘‘MeanAudio: Fast and faithful text-to-audio generation with mean flows,’’ *arXiv preprint arXiv:2508.06098*, 2025.
- [14] Hyung Won Chung et al., ‘‘Scaling instruction-finetuned language models,’’ *Journal of Machine Learning Research*, 2024.
- [15] Jin Xu et al., ‘‘Qwen3-Omni technical report,’’ *arXiv preprint arXiv:2509.17765*, 2025.
- [16] Arijit Biswas and Lars Villemoes, ‘‘Towards evaluating generative audio: Insights from neural audio codec embedding distances,’’ *arXiv preprint arXiv:2509.18823*, 2025.
- [17] Florian Grötschla et al., ‘‘Benchmarking music generation models and metrics via human preference studies,’’ in *ICASSP*, 2025.
- [18] Weiwei Li, ‘‘MeanAudio-S with DACO: Efficient text-to-music generation via rectified flow and distribution-aware posterior refinement,’’ in *ICME Grand Challenge Paper*, 2026.
- [19] Anthony Wang and Shlomo Dubnov, ‘‘Efficient text-to-music generation via flow matching with bidirectional mamba SSM,’’ in *ICME Grand Challenge Paper*, 2026.
- [20] Yonghyun Kim et al., ‘‘Improving text-to-music generation with human preference rewards,’’ in *ICME Grand Challenge Paper*, 2026.
- [21] Junyoung Koh, ‘‘Instrumental text-to-music generation with auxiliary conditioning branches,’’ in *ICME Grand Challenge Paper*, 2026.
- [22] Huakang Chen et al., ‘‘S2Accompanist: A semantic-aware and structure-guided diffusion model for music accompaniment generation,’’ in *ICME Grand Challenge Paper*, 2026.
- [23] Yun-Chen Cheng, Tzu-Hung Huang, and Chih-Pin Tan, ‘‘Making the most of limited data: Score-aware training for text-to-music generation,’’ in *ICME Grand Challenge Paper*, 2026.
- [24] Shunsuke Yoshida et al., ‘‘UT-AISTimpr submission for ICME 2026 grand challenge on academic text-to-music generation,’’ in *ICME Grand Challenge Paper*, 2026.
- [25] Yuqing Cheng, Xingyu Ma, Guochen Yu, and Xiaotao Gu, ‘‘Modeling music as a time-frequency image: A 2D tokenizer for music generation,’’ in *ICME Grand Challenge Paper*, 2026.