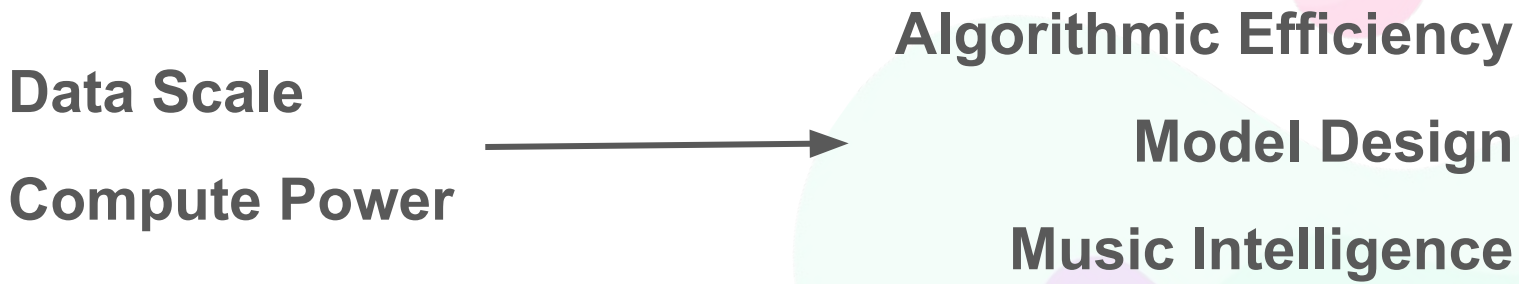
The Academic Text-To-Music Grand Challenge

IEEE ICME 2026

July 9th, 2026 @ Bangkok

Motivation

Existing SOTA Text-to-Music systems rely on proprietary datasets and industrial-scale hardware, creating a steep barrier for academic systems compete against.



Oraganizers



Yi-Hsuan Yang



Hao-Wen Dong



Hung-Yi Lee



Fang-Chih Hsieh



Wei-Jaw Lee

Agenda

Time slot	Subject
9:30 am - 9:35 am	Opening
9:35 am - 9:50 am	Talk 1: Academic Text-to-Music Grand Challenge: Results & Lessons Learned
9:50 am - 10:05 am	Talk 2: UT-AISTimprt submission for ICME 2026 Grand Challenge on Academic Text-to-Music Generation
10:05 am - 10:20 am	Talk 3: Instrumental Text-to-Music Generation with Auxiliary Conditioning Branches
10:20 am - 10:35 am	Talk 4: MeanAudio-S with DACO: Efficient Text-to-Music Generation via Rectified Flow and Distribution-Aware Posterior Refinement
10:35 am - 10:50 am	Talk 5: Improving Text-to-Music Generation with Human Preference Rewards
10:50 am - 11:10 am	20-Minute Break

Agenda

Time slot	Subject
11:10 am - 11:25 am	Talk 6: Efficient Text-to-Music Generation via Flow Matching with Bidirectional Mamba SSM
11:25 am - 11:40 am	Talk 7: Making the Most of Limited Data: Score-Aware Training for Text-to-Music Generation
11:40 am - 11:55 am	Talk 8: S2Accompanist: A Semantic-Aware and Structure-Guided Diffusion Model for Music Accompaniment Generation
11:55 am - 12:10 pm	Talk 9: Modeling Music as a Time-Frequency Image: A 2D Tokenizer for Music Generation

Presentation Guidelines

12

Minutes

Presentation

3

Minutes

Q&A

A signal will be shown when there are 2 minutes remaining for the presentation.

ATTM | The Fair-Play Constrains

Task

Generate 10-second instrumental music clip from text prompt.

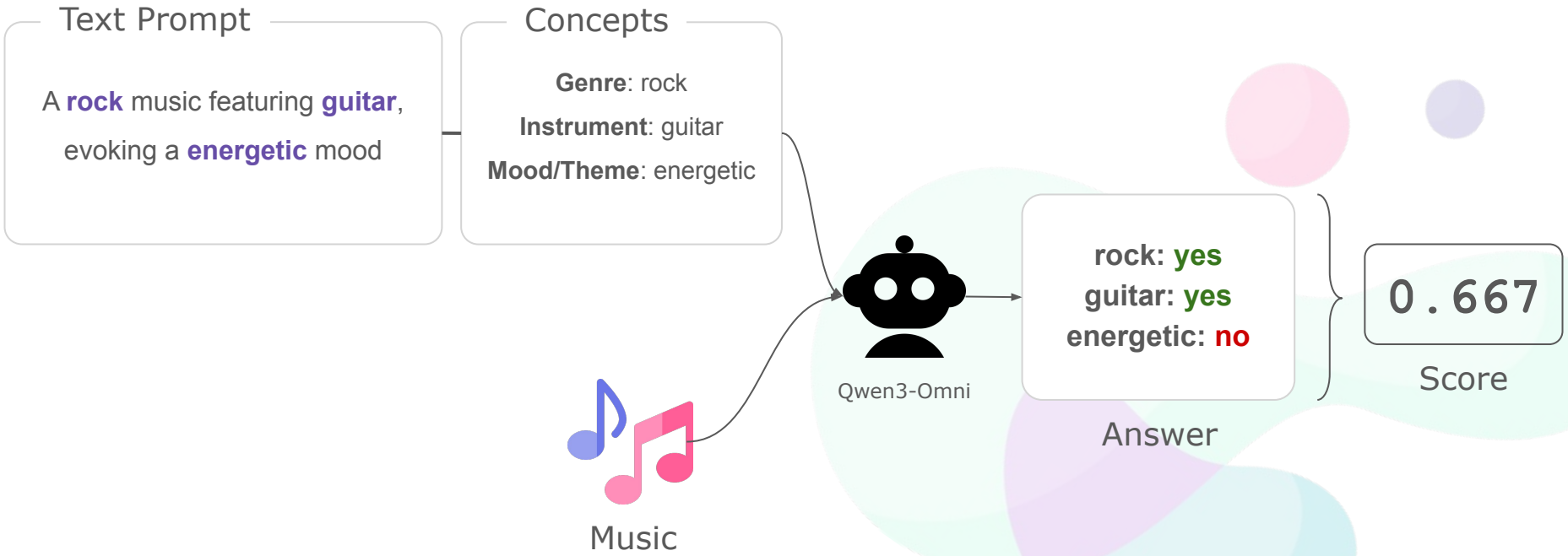
From-Scratch Training

Core generative model must be trained from-scratch.

Data Restriction

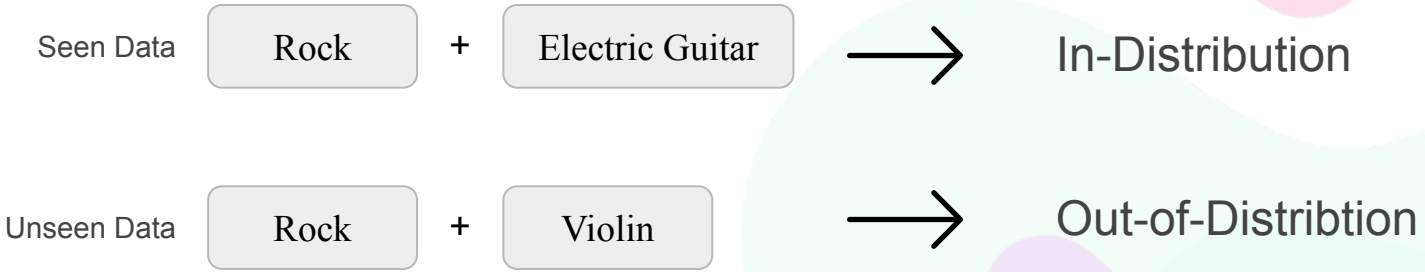
Exclusive use of MTG-Jamendo dataset for all training stages.

ATTM | Concept Coverage Score (CCS)



ATTM | The Out-of-Distribution Problem

We curated In-Distribution (ID) and Out-of-Distribution (OOD) prompt sets to analyze compositional performance.



Participants Overview

Submission ID(s)	Institution
e00, p00	Tencent, China
e01	UCSD, United States
e02	CMU, Sony AI, Georgia Tech, Peking University, KAIST, QMUL, and Sony Group Corporation; multi-country collaboration
e03	Hiroshima University, Japan
e04	Hiroshima University, Japan
e05, p05	Yonsei University, South Korea
e06	Munich Music Labs, Germany
e07	Northwestern Polytechnical University, China
e08	National Taiwan University, Taiwan
e09, p09	University of Tokyo and AIST, Japan
e10, p10	Central Conservatory of Music, China
e11	CMU, National Taiwan University, and University of Washington; United States and Taiwan

Objective Results

Model	Params	Train Data (hours)	Arch.	A100-eq GPU-H	FAD↓	CLAP↑	CCS↑	Rank
e00	120M	0.46K	D/F, T	48.0	0.556	0.310	0.796	6
e01	189M	3.7K	D/F, SSM	25.1	0.577	**0.338**	0.863	2
e02	120M	3.7K	D/F, T	14.2	0.498	0.270	0.763	8
e03	340M	0.46K	D/F	127.0	0.518	0.251	0.763	12
e04	70M	0.46K	D/F, T	3.8	0.574	0.195	0.833	9
e05	499M	0.46K	D/F, T	20.0	0.487	0.305	0.800	2
e06	488M	0.46K	T, SSM	38.1	0.667	0.268	0.808	9
e07	402M	3.7K	D/F	148.9	**0.417**	0.261	**0.867**	1
e08	450M	3.7K	D/F, T	190.3	0.495	0.295	0.804	2
e09	480M	3.7K	D/F, T	168.1	0.646	0.263	0.767	12
e10	315M	0.46K	T	482.1	0.482	0.163	0.738	11
e11	480M	0.46K	D/F	84.6	0.892	0.097	0.675	16
p00	502M	0.46K	D/F, T	48.0	0.557	**0.311**	0.796	6
p05	2.4B	0.46K	D/F, T	60.0	0.514	0.306	**0.800**	5
p09	480M	3.7K	D/F, T	168.1	0.646	0.260	0.767	15
p10	1.5B	0.46K	T	1319.3	**0.500**	0.171	0.721	14
FluxAudio-S (Baseline)	120M	3.7K	D/F, T	25.8	0.757	0.088	0.592	17
Stable Audio Open	1.1B	7.3K	D/F	N/A	0.574	0.321	0.800	–
MusicGen-small	300M	20K	T	N/A	0.574	0.370	0.875	–
MusicGen-medium	1.5B	20K	T	N/A	0.548	0.353	0.892	–
MusicGen-large	3.3B	20K	T	N/A	0.553	0.379	0.888	–
MeanAudio-S-Full	120M	10K	D/F, T	N/A	0.649	0.210	0.808	–
MeanAudio-L-Full	480M	10K	D/F, T	N/A	0.660	0.202	0.783	–

A100_equivalent_GPU_hours = reported_GPU_hours * (GPU_dense_FP16_TFLOPS / 312)

Subjective Results

35 Listeners

25 Experts

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

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

Lessons Learned from ATTM #1

Data Quality & Supervision Design Are Critical

- **e07** rebuilt the data pipeline rather than only scaling training.
  - Re-separated audio, performed structural segmentation, generated segment-level captions, and applied dual-metric quality grading.
  - Result: strong objective ranking and Efficiency Track 1st.
- **e08** treated alignment as a training signal.
  - Used CLAP-based filtering, CLAP-aware Beta scheduling, and caption style rewriting.
  - Key idea: low-alignment data should be handled differently, not simply discarded.

e07: S2Accompanist: A Semantic-Aware and Structure-Guided Diffusion Model for Music Accompaniment Generation
e08: Making the Most of Limited Data: Score-Aware Training for Text-to-Music Generation

Lessons Learned from ATTM #2

Quality-Aware Post-Training Is An Effective Trick

- **e07** used a two-stage training strategy.
  - First trained on the full corpus to build a broad generative prior.
  - Then fine-tuned on the top-quality subset selected by structure and semantic quality scores.
- **e08** improved training through quality-aware scheduling.
  - Used CLAP alignment scores for filtering and noise timestep scheduling.
  - Also used caption rewriting to bridge verbose training captions and concise inference prompts.
- **e02** explored expert iteration with preference signals.
  - Generated candidate samples, ranked them with TuneJury reward and CLAP-text similarity.
  - Fine-tuned on top-decile samples; internal ablation identified expert iteration as the dominant gain.

e07: S2Accompanist: A Semantic-Aware and Structure-Guided Diffusion Model for Music Accompaniment Generation
e08: Making the Most of Limited Data: Score-Aware Training for Text-to-Music Generation
e02: Improving Text-to-Music Generation with Human Preference Rewards

Lessons Learned from ATTM #3

Auxiliary Components Can Make a Huge Difference

- **e01** showed that a strong audio latent space enables efficient models.
  - Used ACE-Step autoencoder with a 189M Mamba-SSM flow-matching model.
  - Achieved Efficiency Track 2nd despite a much smaller core model.
- **e08** used pretrained representations as auxiliary supervision.
  - Used ACE-Step encoder in the generation pipeline.
  - Applied CLAP and MuQ alignment to transfer semantic and music-structural knowledge.
- **e07** directly improved the audio representation bottleneck.
  - Proposed semantic-aware VAE fine-tuning.
  - Improved the VAE latent space to better capture musical structure, not just audio reconstruction.

e01: Efficient Text-to-Music Generation via Flow Matching with Bidirectional Mamba SSM
e08: Making the Most of Limited Data: Score-Aware Training for Text-to-Music Generation
e07: S2Accompanist: A Semantic-Aware and Structure-Guided Diffusion Model for Music Accompaniment Generation

Lessons Learned from ATTM #4

Diffusion / Flow Matching Dominate; SSM Is A Promising Efficient Path

- **Diffusion / Flow Matching dominated the strongest submissions.**
  - e07, e05/p05, e08, e01, e02, and e09/p09 all used diffusion or flow-matching-based systems.
  - Finalist systems were overwhelmingly from this paradigm.
- **SSM offers a promising efficiency direction.**
  - e01 combined bidirectional Mamba SSM with flow matching.
  - A 189M model achieved Efficiency Track 2nd, suggesting SSMs may be a strong alternative to attention-heavy Transformers.

e07: S2Accompanist: A Semantic-Aware and Structure-Guided Diffusion Model for Music Accompaniment Generation
e05/p05: Instrumental Text-to-Music Generation with Auxiliary Conditioning Branches
e08: Making the Most of Limited Data: Score-Aware Training for Text-to-Music Generation
e01: Efficient Text-to-Music Generation via Flow Matching with Bidirectional Mamba SSM
e02: Improving Text-to-Music Generation with Human Preference Rewards
e09/p09: UT-AISTimprt submission for ICME 2026 Grand Challenge on Academic Text-to-Music Generation