EchoScript: Enhancing AI Music Generation for Cinematic Scoring via Script-Aware Fine-Tuning

Mohammad Kasra Sartaee, Kayvan Karim

Heriot-Watt University, Department of Computer Science
kasrasartaee@yahoo.com, k.karim@hw.ac.uk

Abstract

Recent advancements in artificial intelligence (AI) have significantly transformed the landscape of music generation, enabling context-sensitive and emotionally expressive soundtracks for diverse media applications such as film, gaming, and therapeutic environments. However, existing AI models continue to face persistent challenges in maintaining melodic coherence, thematic continuity, and emotional depth—qualities essential for professional soundtrack production.

This research addresses these limitations by fine-tuning **MusicGen**, a transformer-based generative AI model, to create **EchoScript**—an optimized variant specifically tailored for cinematic soundtrack composition through script-driven conditioning. A curated dataset enriched with detailed metadata, including genre, mood, instrumentation, tempo, and narrative context, was employed to guide the fine-tuning process.

Evaluation results demonstrate substantial improvements over the baseline model. EchoScript achieved a lower **Fréchet Audio Distance (FAD)** score (**4.3738** vs. **4.5492**) and outperformed the baseline in structured listening tests, with participants consistently preferring EchoScript for musical quality and narrative alignment.

Beyond these empirical findings, the study critically examines technical constraints and outlines key future directions, including symbolic-audio integration, enhanced audio mixing, and the development of standardised evaluation metrics. Collectively, these contributions advance the pursuit of AI-generated music that closely approximates human-level expressiveness and narrative coherence, offering meaningful benefits for creative industries reliant on adaptive and emotionally resonant soundtracks.

Introduction

Artificial Intelligence (AI) has rapidly transformed the landscape of music generation, offering new possibilities for composers, developers, and media creators. In dynamic contexts such as film, television, and video games, there is increasing demand for adaptive soundtracks that not only complement emotional tone but respond fluidly to narrative developments. Music in these settings is not merely decorative; it shapes emotional experiences, builds tension, and reinforces storytelling.

Despite notable progress, current AI models still face critical limitations in professional soundtrack composition. Existing systems often lack the structural coherence, emotional nuance, and narrative alignment required for high-quality production. Models such as **MusicGen** demonstrate competence in genre transfer and stylistic variety, yet frequently fall short in sustaining long-term melodic progression and adapting meaningfully to narrative prompts. Moreover, evaluating AI-generated music remains a persistent challenge, constrained by limited objective metrics and subjective variability in listener preferences.

This research investigates whether fine-tuning **MusicGen** can improve its narrative responsiveness and emotional expressiveness, particularly for soundtrack generation. By curating a specialised dataset of cinematic compositions enriched with metadata — including genre, tempo, key, mood, and scene descriptions — and applying targeted training strategies, the study aims to enhance both the technical fidelity and narrative relevance of generated outputs.

This work makes the following contributions:

1. **Model Fine-Tuning:** Demonstrates how task-specific fine-tuning enhances MusicGen's ability to generate emotionally and narratively aligned soundtracks.
2. **Dataset Development:** Presents a curated, soundtrack-focused dataset with detailed metadata annotations to guide the generation process.
3. **Evaluation Framework:** Implements a dual evaluation approach, combining *Fréchet Audio Distance (FAD)* as an objective metric with structured human listening tests for subjective assessment.
4. **Analysis of Model Limitations:** Identifies ongoing challenges, including computational constraints, evaluation difficulties, and the absence of integrated symbolic controls, providing insights for future research.

The findings indicate that fine-tuning substantially improves the musical coherence, emotional depth, and narrative relevance of generated compositions. However, challenges remain in areas such as audio mixing and mastering quality, scalability, and the integration of symbolic control mechanisms for more precise compositional direction.

Background

AI-based music generation has evolved significantly from its early rule-based systems to advanced deep learning models capable of producing expressive, coherent, and stylistically complex compositions. Historically, music composition required extensive expertise in theory, emotion, and technical execution. Initial attempts to automate this process were predominantly rule-based systems, capable of producing simple musical patterns but lacking the emotional depth and structural sophistication needed for professional contexts (Creswell et al. 2018; Pathariya et al. 2024).

The introduction of machine learning, and in particular deep learning, marked a major turning point in the field. These approaches enabled models to learn from large datasets, capturing intricate patterns in melody, harmony, and emotion. Techniques such as **Recurrent Neural Networks (RNNs)** and **Long Short-Term Memory (LSTM)** models proved effective in handling sequential data, capturing temporal dependencies critical to music generation (Briot, Hadjeres, and Pachet 2017; Pathariya et al. 2024). However, these architectures struggled with longer-term dependencies and complex musical structures.

To address these challenges, **Variational Autoencoders (VAEs)** were developed to interpolate between musical styles within a learned latent space, enabling smoother transitions and greater diversity in generated outputs (Pathariya et al. 2024). Despite these advancements, VAEs continued to struggle with maintaining coherence over longer compositions.

Generative Adversarial Networks (GANs) further advanced the field by employing adversarial training to enhance realism and polyphony (Dong et al. 2018a). MuseGAN, a notable GAN-based model, focused on multi-track generation, producing coherent polyphonic compositions across multiple instrument layers. Nevertheless, GANs presented issues such as training instability and mode collapse, limiting the diversity of outputs (Pathariya et al. 2024).

Transformer-based architectures, exemplified by Music Transformer and OpenAI's Jukebox, introduced self-attention mechanisms that significantly improved the handling of long-range dependencies and hierarchical musical structures (Copet et al. 2023; Mittal et al. 2021). MusicGen and Jukebox effectively demonstrated high-quality generation of coherent, expressive music, leveraging text or melodic conditioning to guide outputs (Mittal et al. 2021).

Most recently, **diffusion models** such as MusicLDM and Moûsai have emerged, generating music by iteratively refining noise into structured audio (Schneider et al. 2024; Mittal et al. 2021). These models achieve high fidelity, complex musical textures, and adaptability to prompts, proving particularly effective for real-time applications in gaming and film scoring.

Alongside these technical advancements, both open-source and commercial models have flourished. **MusicGen** leverages transformer architecture to generate harmonically rich music conditioned on text inputs, offering creators considerable control (Copet et al. 2023; Koo et al. 2024). **Moûsai**, using diffusion-based techniques, produces nuanced long-form compositions with high fidelity (Schneider et al. 2024). **MuseGAN** excels in multi-track generation, while proprietary models like Jukedeck and OpenAI's Jukebox have demonstrated commercial success, albeit with accessibility limitations due to closed-source restrictions (Dong et al. 2018a; Briot, Hadjeres, and Pachet 2017).

Evaluation Techniques in AI Music Generation

Evaluating AI-generated music remains a multifaceted challenge. Objective metrics such as **Fréchet Audio Distance (FAD)** quantitatively assess realism by comparing distributions of generated and real audio samples (Kilgour et al. 2019). The **Inception Score (IS)** evaluates clarity and diversity in outputs (Barratt and Sharma 2018), while **Kullback-Leibler (KL)** divergence measures alignment with target emotional tones (Raiber and Kurland 2017).

Additional metrics, including **pitch and rhythm consistency** (Yang and Lerch 2020), **self-similarity analysis** (Dervakos, Filandrianos, and Stamou 2020), and **latent space interpolation** (Shlens 2014), help ensure internal coherence and stylistic variation.

Subjective listening tests remain indispensable, capturing human perception of aesthetic appeal, genre authenticity, and emotional expressiveness. Participants in these tests assess musicality, genre accuracy, emotional alignment, and structural coherence (Schneider et al. 2024; Mittal et al. 2021; Dong et al. 2018a; Copet et al. 2023).

Task-specific evaluations, such as **track matching** in multi-track generation models like MuseGAN (Dong et al. 2018a), and relevance metrics like the **CLAP score** for text-to-music models like MusicGen (Copet et al. 2023), further enrich the evaluation framework. In rhythm-focused applications, **beat synchronisation metrics** ensure timing accuracy, crucial for contexts like dance or video game scoring (Mittal et al. 2021).

Evaluation also extends to model-specific considerations. GANs, for example, require monitoring for **mode collapse** and training instability (Pathariya et al. 2024; Dong et al. 2018a). Transformer models are evaluated for **generalization** to unseen data through cross-validation (Mittal et al. 2021).

Real-world evaluations play an essential role in assessing practical applicability. In gaming and film, **adaptability to narrative shifts** is critical (Pathariya et al. 2024), while in therapeutic contexts, AI-generated music is evaluated for its emotional impact, using both physiological indicators and subjective feedback (Schneider et al. 2024).

Summary of Progress

Overall, the evolution of AI music generation reflects a continual drive toward models that balance creativity, control, and efficiency. Building on this foundation, the present research aims to advance AI-generated soundtracks by enhancing narrative alignment and emotional depth through targeted fine-tuning strategies.

Model and Architecture		Dataset & Representation	Loss Function	Focus / Use Case
CNN-Based Models				
WaveNet (2016)	CNN	VCTK, YouTube Data (Waveform)	L1 Loss	High-fidelity speech synthesis
DCGAN (2016)	CNN	Lakh MIDI (Waveform)	Binary Cross-Entropy	Polyphonic music generation
LSTM-Based Models				
BachBot (2016)	LSTM	Bach Chorale (Symbolic Data)	Cross-Entropy Loss	Harmonization of Baroque music
DeepBach (2017)	LSTM	Bach Chorale (MIDI File)	Cross-Entropy Loss	Harmonization, structured sequences
GAN-Based Models				
MuseGAN (2018)	GAN	Lakh MIDI (Multi-track MIDI)	Binary Cross-Entropy	Multi-instrument polyphony
WaveGAN (2019)	GAN	Speech Commands, AudioSet (Waveform)	Wasserstein Distance	Realistic audio waveform generation
Transformer-Based Models				
Music Transformer (2019)	Transformer	Lakh MIDI (MIDI File)	Cross-Entropy Loss	Melody and harmony generation
MusicLM (2023)	Transformer + AudioLDM	Free Music Archive (Waveform)	Cross-Entropy, Contrastive Loss	Text-to-music generation with thematic control
MusicGen (2023)	Transformer	Shutterstock, Pond5 (Waveform)	Cross-Entropy, Perceptual Loss	High-fidelity music from text prompts
Diffusion Models				
DiffWave (2020)	Diffusion Model	VCTK, LJSpeech (Waveform)	L1 Loss, GAN Loss	High-quality synthesis of waveforms
MusicLDM (2023)	Diffusion Model	Moûsai-2023 (Mel-spectrogram)	Diffusion Loss	High-fidelity adaptive audio
Noise2Music (2023)	Diffusion Model	MusicCaps, MTAT (Waveform)	Diffusion Loss	Text-conditioned music composition

Table 1: Chronological Overview of Representative Music Generation Models with Key Technical Details, adapted from (Chen, Huang, and Gou 2024).

Methodology

This research followed a multi-stage methodology encompassing dataset curation, extensive preprocessing, model fine-tuning, and exploratory investigation of hybrid generation strategies. The objective was to build a fully customised pipeline for soundtrack generation capable of producing emotionally expressive, narratively aligned compositions, while laying the groundwork for future integration of symbolic controls.

Dataset Development and Preprocessing

Due to the absence of open-source, soundtrack-specific datasets, a custom dataset was created from scratch, designed to simulate real-world soundtrack generation requirements.

Data Acquisition and Organization High-quality, copyright-free cinematic music tracks were sourced from verified repositories. Data management was centralized via **Google Drive**, which facilitated clear organization into distinct directories for training, validation, and evaluation. This structure also enabled scalability for future dataset expansions.

Feature Extraction and Metadata Annotation The **PANNs** framework extracted detailed acoustic attributes such as genre, mood, and instrumentation, providing a semantic foundation for the dataset. Tagging categories were standardized as follows:

- **Genres:** Classical, electronic, orchestral, ambient, cinematic.
- **Moods:** Uplifting, suspenseful, melancholic, calming.
- **Instrumentation:** Guitar, piano, violin, synthesizer, percussion.

Further contextual enrichment was performed using the **GPT-4 API**, which generated approximately 170 diverse scene-based prompts and keywords. These text descriptions were carefully engineered to enhance the multimodal pairing between narrative intent and musical output.

Multimodal Alignment Using CLAP The **CLAP** model was leveraged to measure semantic similarity between audio samples and their corresponding textual descriptions. This ensured high-fidelity alignment between metadata and audio content, with refined prompt selection based on embedding closeness.

Audio Standardisation and Feature Engineering All audio files were uniformly resampled to **44.1 kHz** and segmented into consistent **30-second clips** to streamline training input. Acoustic features including **MFCCs**, **Mel Spectrograms**, **Chroma Features**, and global parameters such as tempo and key signature (via **Librosa**) were extracted to support both data analysis and potential auxiliary training signals.

Metadata were consolidated into a comprehensive **JSON schema**, recording song-level descriptors, file paths, genre tags, mood labels, instrumentation, keywords, and scene

narratives. This ensured compatibility with downstream pipelines and reproducibility of experiments.

Model Selection and Fine-Tuning Strategy

MusicGen was selected for its transformer-based autoregressive architecture, known for generating coherent and high-fidelity musical outputs. Its single-stage pipeline offered superior computational efficiency over diffusion models, with pretrained capabilities across a broad musical spectrum.

The **small variant** of MusicGen was chosen to balance computational feasibility with output quality, given the resource constraints of the training environment.

Customisation for Soundtrack Composition
The fine-tuning process involved:

- **Training on the custom dataset** enriched with detailed semantic annotations and scene narratives.
- Incorporating **melody conditioning** via chromagrams to enhance narrative responsiveness and harmonic structure.
- Applying **regularization techniques**, including dropout layers and Classifier-Free Guidance (CFG), to mitigate overfitting.
- Utilizing **weighted sampling** to balance genre representation and avoid bias toward dominant categories.

Training was executed on **NVIDIA A100 GPUs** (40 GB VRAM), utilizing **FP16 mixed precision** and **FlashAttention** to optimize memory usage and learning efficiency.

Training Configuration

The final training configuration included:

- **Batch Size:** 12
- **Learning Rate:** 5×10^{-5} with cosine annealing scheduler
- **Epochs:** 10
- **Optimizer:** AdamW with weight decay
- **Loss Function:** Categorical cross-entropy for token prediction
- **Gradient Accumulation:** Enabled to efficiently utilize GPU memory

Training encompassed approximately **5,000 audio samples**, with systematic checkpointing after each epoch for iterative refinement.

Exploration of Hybrid MIDI-Audio Generation

This study initially aimed to integrate symbolic control using Music ControlNet, but its proprietary nature and lack of open-source access made integration infeasible. Alternative approaches like MIDI-based annotation and audio-to-MIDI alignment were explored, though they offered limited effectiveness. Future research could revisit hybrid strategies as symbolic-audio frameworks become more accessible.

Evaluation

To comprehensively assess the performance of the fine-tuned **EchoScript** model in soundtrack generation, this study employed a dual evaluation strategy: one quantitative, using objective audio fidelity metrics, and the other qualitative, involving structured human listening tests. This combined approach was designed to evaluate both the technical realism of the outputs and their narrative alignment and emotional relevance, thereby addressing the multifaceted nature of music evaluation. A comparative assessment against the baseline **MusicGen** model was conducted under identical conditions to isolate the effect of fine-tuning.

Objective Evaluation Protocol

Objective evaluation metrics are essential for ensuring consistency and reproducibility in generative audio research. For this study, the primary metric selected was the **Fréchet Audio Distance (FAD)**, which is widely recognized for its ability to capture perceptual audio quality by comparing statistical properties of real and generated music.

Fréchet Audio Distance (FAD) The **Fréchet Audio Distance (FAD)** metric, introduced by Kilgour et al. (2019), extends the logic of Fréchet Inception Distance (FID) from image to audio domains. It quantifies the statistical distance between the distributions of feature embeddings extracted from real background audio and generated samples, using a pretrained audio classification model (e.g., VGGish or YamNet).

FAD treats each set of audio embeddings as multivariate Gaussian distributions and computes their distance using the following formula:

$$\mathrm{FAD}(N_b, N_e) = \|\mu_b - \mu_e\|^2 + \mathrm{tr}(\Sigma_b + \Sigma_e - 2\sqrt{\Sigma_b \Sigma_e}) \quad (1)$$

where:

- μ_b, Σ_b: Mean and covariance of embeddings from background (real) audio tracks.
- μ_e, Σ_e: Mean and covariance of embeddings from evaluation (generated) audio tracks.
- $\mathrm{tr}(\cdot)$: Trace of the matrix.

Lower FAD scores imply that the generated audio more closely resembles professionally composed music in the perceptual feature space. For consistency, a high-quality background dataset of professionally composed, license-compliant cinematic music was curated and used as the reference distribution against which both EchoScript and MusicGen were evaluated.

To minimise experimental variability, all generated samples were preprocessed with consistent sampling rates, durations (30 seconds), and formatting. Each model's output was embedded and analysed independently using the same pretrained classifier to ensure consistency in the FAD computation pipeline.

Subjective Listening Test Design

Objective metrics are valuable but insufficient for capturing the artistic and perceptual dimensions of music. Therefore,

a structured human listening study was conducted to assess EchoScript's performance in terms of musical quality and narrative alignment. This component aimed to approximate real-world listener perception in media and entertainment contexts.

Participant Profile and Selection Twenty participants were recruited for the study through purposive sampling. The sample mostly included young adults aged 18–24, representing a mix of musical backgrounds, from casual listeners to hobbyist musicians. A demographic survey administered before the listening session captured information on:

- Age, gender, and education level.
- Musical training and experience.
- Familiarity with AI-generated music.
- Listening habits (e.g., daily engagement with instrumental or cinematic music).

All participants provided informed consent and were briefed about the nature of the study, ensuring ethical compliance and voluntary participation.

Listening Task Design Participants listened to a total of ten audio clips — five generated by EchoScript and five by the baseline MusicGen — each paired to a specific script-based scenario. Scenarios were carefully selected to reflect diverse emotional and narrative contexts typical of soundtrack applications (e.g., suspense, tranquillity, action, melancholy, and triumph). For each scenario, participants listened to two anonymised, randomised samples (one from each model) without knowledge of which system generated which clip.

Evaluation Criteria and Scoring Method After listening to each pair, participants completed a structured Likert-scale questionnaire across two major dimensions:

- **Musical Quality (MOS)** – Measures technical and aesthetic qualities including:
  - Clarity and absence of distortion.
  - Professionalism and production value.
  - Naturalness of transitions and musical phrasing.
  - Balance and coherence of instrumentation.
- **Relevance to Script (REL)** – Measures contextual alignment including:
  - Mood appropriateness and emotional expressiveness.
  - Narrative scene alignment and thematic fit.
  - Responsiveness to detailed script cues (e.g., tempo, dynamics).
  - Overall suitability for use in film/game context.

Each sub-question was scored on a 5-point Likert scale (1 = very poor, 5 = excellent). Additionally, participants indicated a binary preference for each pair, selecting the sample they felt better matched the scenario.

Environment and Data Handling All listening tests were conducted in quiet environments using high-quality over-ear headphones to minimise distractions and ensure audio fidelity. Responses were collected digitally via a structured form and anonymised before analysis. The quantitative scores were processed into aggregate MOS and REL ratings per model per scenario, while binary preferences were tallied across participants to determine model preference distributions.

Evaluation Scope and Coverage

This evaluation methodology was designed to offer broad yet rigorous coverage of:

- **Technical fidelity** – via objective metrics like FAD.
- **Perceptual coherence and quality** – via human-assessed MOS.
- **Narrative and emotional alignment** – via REL ratings and preference choices.

By combining automated and human-centred evaluation techniques, the methodology supports a holistic assessment of the model's fitness for real-world soundtrack applications. All evaluation protocols were executed in a consistent, replicable manner to ensure fairness across model comparisons.

The following section presents the results and insights derived from this comprehensive evaluation.

Results

This section presents the outcomes of the evaluation, reporting both objective metrics and subjective listening test results to assess the performance of the fine-tuned **EchoScript** model compared to the baseline **MusicGen** model. The results demonstrate EchoScript's superiority in terms of audio realism, musical quality, and narrative alignment.

Objective Evaluation Results

The primary objective metric used in this study was the **Fréchet Audio Distance (FAD)**, selected for its effectiveness in quantifying audio realism and fidelity.

EchoScript achieved an FAD score of **4.3738**, compared to the baseline MusicGen's score of **4.5492**. This measurable improvement indicates that fine-tuning successfully enhanced the model's ability to capture detailed acoustic characteristics and produce audio more closely aligned with professional-quality soundtracks.

Model	FAD Score ↓
EchoScript (Fine-tuned)	**4.3738**
MusicGen (Baseline)	4.5492

Table 2: Fréchet Audio Distance (FAD) Comparison

These results empirically validate the effectiveness of the adopted fine-tuning strategy and underscore EchoScript's potential for use in immersive media applications.

Subjective Listening Test Results

A structured listening test involving twenty participants was conducted to evaluate EchoScript against MusicGen across two key dimensions: **Musical Quality (MOS)** and **Relevance to Script (REL)**.

Participants, primarily young adults aged 18–24 with diverse educational backgrounds and varying degrees of musical experience, evaluated five distinct script-based scenarios.

Musical Quality (MOS) EchoScript consistently outperformed MusicGen across all intrinsic musical quality criteria:

Attribute	EchoScript	MusicGen
Audio Clarity	4.02	3.64
Production Quality	4.08	3.54
Transition Smoothness	4.25	3.48
Instrument Balance	3.95	3.64

Table 3: Mean Opinion Scores (MOS) for Musical Quality

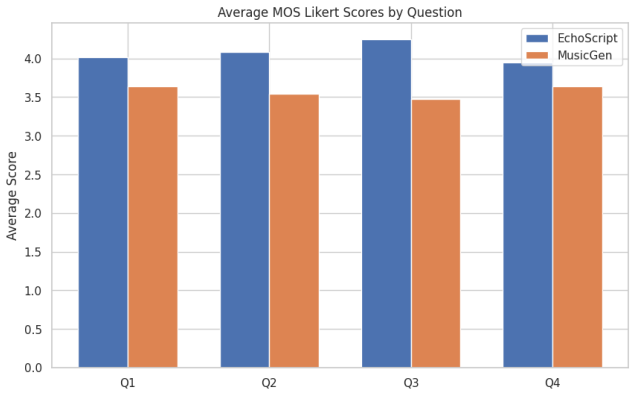

Figure 1: Average MOS Likert scores by question for EchoScript and MusicGen, demonstrating EchoScript's consistent advantage in musical clarity and transitions.

Aggregate MOS ratings further reinforced EchoScript's superiority, with an overall average of **7.66** compared to MusicGen's **6.75**. Moreover, **53.8%** of participants explicitly preferred EchoScript for musical quality.

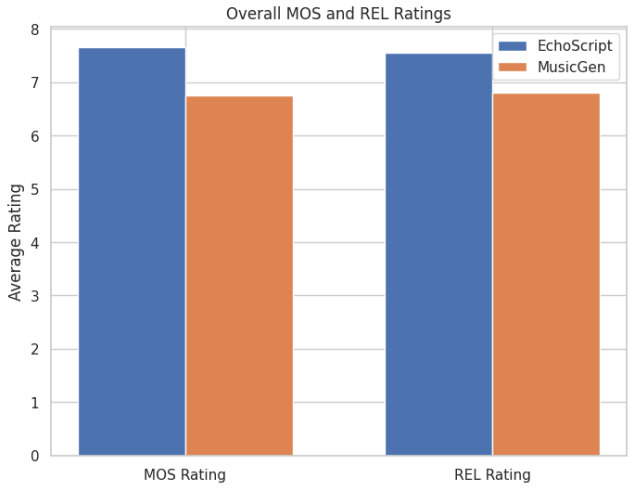

Figure 2: Overall MOS and REL ratings for EchoScript and MusicGen, highlighting EchoScript's higher aggregated performance.

To further visualise rating distributions, the following boxplots illustrate participant feedback spread for MOS and REL ratings.

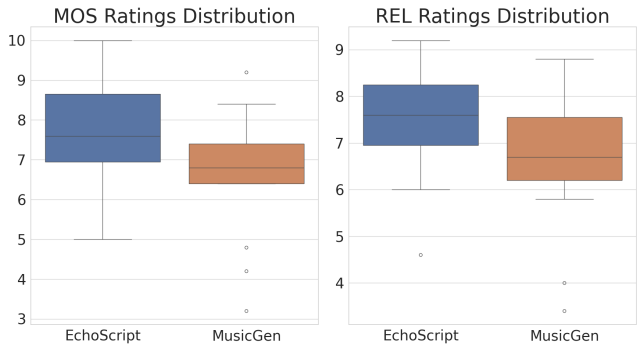

Figure 3: Distribution of MOS and REL ratings across participants, indicating greater consistency and higher scores for EchoScript.

Additionally, explicit participant preferences for MOS are visualised in the following pie chart.

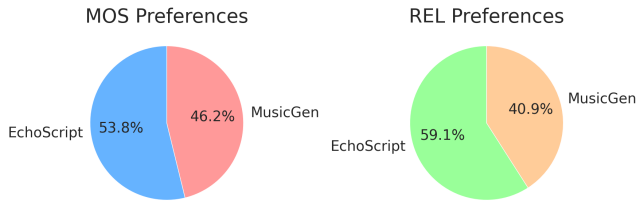

Figure 4: Participant preferences for MOS and REL, with EchoScript favoured in both categories.

Relevance to Script (REL) EchoScript also demonstrated significant improvements in narrative alignment:

Attribute	EchoScript	MusicGen
Mood Conveyance	3.96	3.59
Scene Matching	4.03	3.66
Emotional Expressivity	3.84	3.54
Script Adherence	3.90	3.54

Table 4: Relevance to Script (REL) Scores

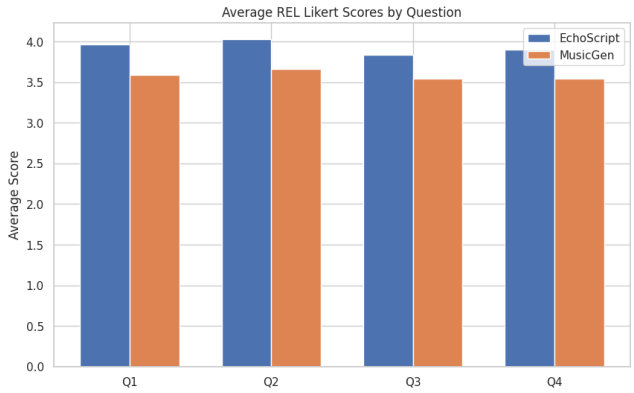

Figure 5: Average REL Likert scores by question for EchoScript and MusicGen, illustrating EchoScript's superior narrative alignment.

Summary of Results

The combined objective and subjective results confirm the superior performance of EchoScript in generating soundtracks that are technically refined, musically coherent, and narratively aligned. Key highlights include:

- EchoScript achieved a lower FAD score, reflecting improved audio realism.
- Subjective evaluations indicated enhanced clarity, production quality, and natural transitions.
- EchoScript demonstrated better alignment with narrative prompts, emotional cues, and detailed script elements.

While EchoScript's results establish a strong foundation, they also highlight areas for future enhancement, particularly in instrumental balance, dynamic range, and emotional depth. These findings lay the groundwork for iterative improvements and further research into fine-tuned, script-conditioned soundtrack generation.

Discussion

This study explored the fine-tuning of the **MusicGen** model for adaptive soundtrack generation, emphasising improvements in emotional expressiveness and narrative alignment through script-based training. While initial ambitions aimed to integrate symbolic music representations with generative audio models, practical constraints necessitated a refined focus on audio-based fine-tuning. Both objective evaluation using **Fréchet Audio Distance (FAD)** and subjective listening tests clearly demonstrated the superiority of the fine-tuned **EchoScript** model in musical coherence, audio clarity, and narrative relevance, affirming the practical effectiveness of the adopted methodology.

Interpreting Results in Context

Quantitative evaluation via FAD revealed substantial improvements, with EchoScript achieving a notably lower score (**4.3738**) compared to the baseline MusicGen (**4.5492**). These findings align with prior studies highlighting the efficacy of tailored training for enhancing acoustic realism and stylistic coherence in generative music systems (Copet et al. 2023; Mittal et al. 2021).

Subjective evaluations further corroborated these insights. Listener preferences strongly favoured EchoScript, with marked improvements in audio clarity, smoothness of transitions, and production quality. These results reinforce existing literature regarding the strengths of transformer-based generative models in achieving structural coherence (Schneider et al. 2024). However, persistent challenges in emotional expressiveness and instrument-level mixing were observed, consistent with earlier research that underscores the difficulty of achieving nuanced emotional depth and detailed sonic balance in AI-generated compositions (Dong et al. 2018a).

Practical Integration Outlook

While this study focused on modelling and evaluation, EchoScript is designed for future integration into real-world media workflows. It could serve as a backend module in DAWs, middleware, or engines like Unity and Unreal, supporting narrative-aligned soundtrack generation. Its structured metadata input makes it suitable for adaptation into plugins or script-driven tools for semi-automated composition.

Conclusion

This study investigated the fine-tuning of the **MusicGen** model for adaptive soundtrack generation, with a focused aim to enhance emotional expressiveness, narrative alignment, and technical fidelity through script-based annotations. While the initial ambition encompassed a hybrid symbolic-audio generation approach, practical constraints necessitated a shift toward an audio-only methodology. Leveraging a meticulously curated dataset enriched with detailed metadata—including genre, mood, instrumentation, tempo, and narrative context—the fine-tuned model, **EchoScript**, demonstrated substantial improvements in musical coherence, emotional depth, and contextual relevance when compared to its baseline counterpart.

Contributions

This research makes several meaningful contributions to the field of AI-driven soundtrack composition, specifically targeting narrative and emotional alignment:

- **Curated and richly annotated dataset creation** — Developed a comprehensive, multi-label dataset with

scene narratives, moods, genres, and instrumentation, addressing the scarcity of resources for soundtrack-specific training.

- **Customised fine-tuning pipeline** — Implemented a training framework incorporating advanced techniques such as weighted sampling, regularisation methods, FP16 precision, and FlashAttention to optimise model performance and efficiency.
- **Objective and subjective evaluation framework** — Established a robust evaluation pipeline combining Fréchet Audio Distance (FAD) and comprehensive listener studies (MOS and REL), setting clear performance benchmarks for future research.
- **Empirical validation of narrative alignment improvements** — Demonstrated measurable enhancements in musical quality and narrative alignment over baseline models, supported by both statistical analyses and participant preference data.
- **Exploration of hybrid MIDI-audio strategies** — Investigated the integration of symbolic control via MIDI, documenting technical challenges and outlining future pathways for enhanced compositional control.
- **Foundation for future extensibility** — Developed an open and extensible research pipeline, facilitating ongoing advancements in hybrid audio-symbolic soundtrack generation.

Collectively, these contributions advance the state of AI-generated music for narrative applications, providing actionable methodologies, performance baselines, and insights for future explorations in emotionally expressive soundtrack generation.

Final Remarks

This research underscores both the substantial potential and the persisting challenges inherent in fine-tuning generative AI models for soundtrack generation. The findings provide clear evidence of the improvements achievable through targeted training strategies while emphasizing the need for continued technical, infrastructural, and methodological advancements. By addressing these complexities, future research can bring AI-generated music closer to human-level expressiveness and professional quality, offering significant benefits to industries reliant on dynamic, adaptive, and emotionally resonant soundtracks.

Limitations and Future Work

While this research achieved meaningful advancements in fine-tuning generative AI models for adaptive soundtrack composition, several limitations constrained the scope and outcomes of the study. Recognizing these challenges provides valuable direction for future research and development in AI-driven music generation.

Identified Limitations

Computational Constraints Due to resource limitations, only the smaller **MusicGen** variant was used. While effective, its limited capacity constrained expressive range. Larger models could offer improved generative performance (Mittal et al. 2021).

Dataset and Annotation Quality The custom dataset relied on AI-assisted tagging, which introduced errors in genre, tempo, and mood labels. The lack of professional, human-curated annotations may have impacted fine-tuning accuracy (Copet et al. 2023).

Symbolic-Audio Integration Plans to integrate symbolic control via **Music ControlNet** were blocked by its proprietary status. The broader lack of open-source symbolic-audio frameworks limited hybrid experimentation (Schneider et al. 2024; Mittal et al. 2021).

Evaluation Limitations The study used FAD and human listening tests, which do not fully capture emotional or structural dimensions of music. A lack of standardised, multi-dimensional metrics remains a challenge in evaluating generative music quality.

Future Directions and Recommendations

This study opens several promising directions for future work:

- **Larger-Scale Training:** Training larger MusicGen variants on high-performance hardware could improve expressive depth and compositional complexity.
- **Improved Datasets:** Publicly available, professionally annotated soundtrack datasets—with support from human-AI tagging—would enhance training quality and reproducibility.
- **Symbolic-Audio Integration:** Developing open-source frameworks for symbolic control (e.g., MIDI) can enable precise melodic and harmonic conditioning in audio generation.
- **Better Evaluation Metrics:** Beyond FAD, hybrid evaluation methods combining human and automated feedback are needed to assess emotional fit and narrative alignment more holistically.
- **Enhanced Audio Quality:** Neural mixing and mastering should be improved to close the gap between AI outputs and professional music production.
- **Broader Applications:** EchoScript's narrative-aware generation has potential in audiobooks, podcasts, e-learning, and tourism—offering immersive, context-driven audio experiences.

By systematically addressing these limitations, future research can advance the state of AI-generated music towards greater expressiveness, precision, and narrative alignment. Such progress will not only benefit academic research but also unlock practical applications in film, gaming, and other creative industries that rely on dynamic, emotionally resonant soundtracks. The foundation established by this study offers a promising launchpad for these continued innovations.

Acknowledgements

The authors would like to thank their families and peers for their continuous encouragement and support throughout the course of this research.

References

Aalbers, S.; Fusar-Poli, L.; Freeman, R.; Spreen, M.; Ket, J.; Vink, A.; Maratos, A.; Crawford, M.; Chen, X.-J.; and Gold, C. 2017. Music therapy for depression. *Cochrane Database of Systematic Reviews*, 1(11).

Agarwal, G.; and Om, H. 2021. An efficient supervised framework for music mood recognition using autoencoder-based optimised support vector regression model. *IET Signal Processing*, 15(2): 98–121.

Barratt, S.; and Sharma, R. 2018. A Note on the Inception Score. *arXiv preprint arXiv:1801.01973*.

Briot, J.-P.; Hadjeres, G.; and Pachet, F.-D. 2017. Deep Learning Techniques for Music Generation – A Survey.

Chen, Y.; Huang, L.; and Gou, T. 2024. Applications and Advances of Artificial Intelligence in Music Generation:A Review.

Chu, H.; Kim, J.; Kim, S.; Lim, H.; Lee, H.; Jin, S.; Lee, J.; Kim, T.; and Ko, S. 2022. An empirical study on how people perceive AI-generated music. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 304–314.

Civit, M.; Civit-Masot, J.; Cuadrado, F.; and Escalona, M. J. 2022. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends.

Cope, D. 1996. Experiments in musical intelligence. 12.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and Controllable Music Generation.

Cosma, O. G.; Domuta, C.; Gota, D.; Stan, O.; Fanca, A.; Pop, A.; Valean, H.; and Miclea, L. 2023. Automatic Music Generation Using Machine Learning. In *International Conference on Electrical, Computer and Energy Technologies, ICECET 2023*. Institute of Electrical and Electronics Engineers Inc. ISBN 9798350327816.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative Adversarial Networks: An Overview.

Cífka, O.; Şimşekli, U.; and Richard, G. 2020. Groove2groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2638–2650.

Dash, A.; and Agres, K. 2024. AI-Based Affective Music Generation Systems: A Review of Methods and Challenges. *ACM Computing Surveys*, 56.

Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Deruty, E.; Grachten, M.; Lattner, S.; Nistal, J.; and Aouameur, C. 2022. On the development and practice of AI technology for contemporary popular music production. volume 5, 35–50.

Dervakos, E.; Filandrianos, G.; and Stamou, G. 2020. Heuristics for Evaluation of AI Generated Music. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 9164–9169. IEEE.

Dhariwal, P.; Jun, H.; Payne, C.; Kim, J. W.; Radford, A.; and Sutskever, I. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

Dong, H.-W.; Hsiao, W.-Y.; Yang, L.-C.; and Yang, Y.-H. 2018a. MuseGAN: Multi-Track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. Thirty-Second AAAI Conference on Artificial Intelligence. ArXiv preprint arXiv:1709.06298.

Dong, H.-W.; Hsiao, W.-Y.; Yang, L.-C.; and Yang, Y.-H. 2018b. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Du, Y.; and Mordatch, I. 2019. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems*, volume 32, 6840–6851. Curran Associates.

Engel, J.; Hantrakul, L.; Gu, C.; and Roberts, A. 2020. DDSP: Differentiable digital signal processing.

Engel, J.; Hoffman, M.; and Roberts, A. 2018. Latent constraints: Learning to generate conditionally from unconditional generative models. In *International Conference on Learning Representations*.

Gan, C.; Huang, D.; Chen, P.; et al. 2020. Foley music: Learning to generate music from videos. In *Computer Vision–ECCV 2020*, 758–775. Springer.

Hadjeres, G.; Pachet, F.; and Nielsen, F. 2017. DeepBach: A steerable model for Bach chorales generation. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 1362–1371. PMLR.

Hawthorne, C.; Stasyuk, A.; Roberts, A.; Simon, I.; Huang, C.-Z. A.; Dieleman, S.; Elsen, E.; Engel, J.; and Eck, D. 2018. Enabling factorized piano music modeling and generation with the MAESTRO dataset. *arXiv preprint arXiv:1810.12247*.

Hawthorne, C.; Stasyuk, A.; Roberts, A.; Simon, I.; Huang, C.-Z. A.; Dieleman, S.; Elsen, E.; Engel, J.; and Eck, D. 2019. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 6840–6851. Curran Associates.

Huang, Q.; Jansen, A.; Lee, J.; et al. 2022. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*.

Huang, Z.; et al. 2023. MusicLM: Generating Music from Text. *arXiv preprint arXiv:2301.11325*.

Kilgour, K.; Zuluaga, M.; Roblek, D.; and Sharifi, M. 2019. Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. *arXiv preprint arXiv:1812.08466*.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

Koo, J.; Wichern, G.; Germain, F.; Khurana, S.; and Roux, J. L. 2024. Understanding and Controlling Generative Music Transformers by Probing Individual Attention Heads.

Liang, F. 2016. Bachbot: Automatic composition in the style of Bach chorales. *University of Cambridge*, 8(3.1).

Lin, T.-F.; and Chen, L.-B. 2024. Harmony and algorithm: Exploring the advancements and impacts of AI-generated music. *IEEE Potentials*, 2–9.

Liu, H.; Chen, Z.; Yuan, Y.; et al. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Mittal, G.; Engel, J.; Hawthorne, C.; and Simon, I. 2021. Symbolic Music Generation with Diffusion Models.

Müller, M. 2015. *Fundamentals of music processing: Audio, analysis, algorithms, applications*, volume 5. Springer.

Norman-Haignere, S. V.; Kanwisher, N.; McDermott, J. H.; and Conway, B. R. 2019. Divergence in the functional organization of human and macaque auditory cortex revealed by fMRI responses to harmonic tones. *Nature neuroscience*, 22(7): 1057–1060.

Pathariya, M. J.; Jalkote, P. B.; Patil, A. M.; Sutar, A. A.; and Ghule, R. L. 2024. Tunes by Technology: A Comprehensive Survey of Music Generation Models. 506–512.

Qian, T.; Kaunismaa, J.; and Chung, T. 2022. cMelGAN: An Efficient Conditional Generative Model Based on Mel Spectrograms. arXiv:2205.07319.

Radford, A.; Kim, J. W.; Hallacy, C.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Raiber, F.; and Kurland, O. 2017. Kullback-Leibler Divergence Revisited. In *Proceedings of the 2017 ACM on International Conference on the Theory of Information Retrieval*, 117–120. ACM.

Salehinejad, H.; Sankar, S.; Barfett, J.; Colak, E.; and Valaee, S. 2018. Recent Advances in Recurrent Neural Networks. *arXiv preprint arXiv:1801.01078*.

Sambaragi, L. M.; Naik, P. P.; Kakatkar, N. N.; Chikkamath, S.; Nirmala, S. R.; and Budihal, S. V. 2024. Music Generation: A simplified approach. In *2024 3rd International Conference for Innovation in Technology, INOCON 2024*. Institute of Electrical and Electronics Engineers Inc. ISBN 9798350381931.

Schneider, F.; Kamal, O.; Jin, Z.; and Schölkopf, B. 2024. Moûsai: Efficient Text-to-Music Diffusion Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8050–8068. Association for Computational Linguistics. We open-source our code and dataset at the provided link.

Shlens, J. 2014. Notes on Kullback-Leibler Divergence and Likelihood Theory. *arXiv preprint arXiv:1404.2000*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 2256–2265. PMLR.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 11895–11907.

Touvron, H.; Lavril, T.; Izacard, G.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Veerendranath, V.; Masti, V.; Gupta, U.; Chaudhuri, H.; and Srinivasa, G. 2023. ScripTONES: Sentiment-Conditioned Music Generation for Movie Scripts. In *ACM International Conference Proceeding Series*. Association for Computing Machinery. ISBN 9798400716492.

Yang, L.-C.; Chou, S.-Y.; and Yang, Y.-H. 2017. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*.

Yang, L.-C.; and Lerch, A. 2020. On the Evaluation of Generative Models in Music. *Neural Computing and Applications*, 32: 4773–4784.

Yin, Z.; Reuben, F.; Stepney, S.; and Collins, T. 2023. Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation. *Machine Learning*, 112: 1785–1822.