

PROJECT ECHO

Presented by:

Deeksha Chauhan
Moksh Aggarwal
Shivam Mahendru
Rajarjeswari Premanand

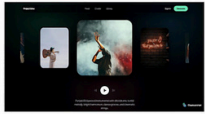
Project Advisors:

Dr. Simon Shim

Introduction

Project Echo is a text-to-music generation system that creates instrumental tracks from natural-language prompts. Built within a strict 15-hour A100 GPU budget, it integrates five model architectures alongside a dual-modality Music RAG module over MusicCaps.

Three core components — M3 (11.3M params), M4 (LoKR-adapted ACE-Step 1.5), and M5 (dual-modal RAG) — unified under a session-safe training pipeline.



- **LoKR adaptation of ACE-Step 1.5** — 0.03% trainable parameters, outperforming LoRA
- **Zero-parameter style transfer** via Attention Intervention and RoPE-enhanced chroma cross-attention,
- **Dual-modal Music-RAG** — caption + CLAP retrieval in <500 ms, human-editable CB0 skeleton

Methodology

Five-Architecture Model Design

Project Echo builds five complementary models across a 15-hour A100 budget, spanning fine-tuning, from-scratch training, parameter-efficient adaptation, and retrieval-augmented generation — all validated with zero decode errors.



- **M1** — MusicGen fine-tuned on 1,161 EDM tracks via chroma cross-attention (~1 hr A100)
- **M2** — ACE-Step + LoKR with style-steerable synthesis via structured text labels (genre, BPM, key, mood)
- **M3** — 11.3M param two-stage transformer trained from scratch; CCS 0.896 across 6 genres in ~40 min
- **M4** — ACE-Step LoKR (0.03% of 4.8B params); SDD ~32.4%, FAD ~23.8% in 15.3 GPU-hours
- **M5** — Caption + CLAP dual-modal retrieval in <500 ms via FastAPI + ChromaDB

Analysis and Results

M4 Primary Evaluation — Base vs. Fine-Tuned

The fine-tuned model achieves consistent improvement across all spectral metrics evaluated on 25 generated samples:

Metric	Base	Fine-Tuned	Delta
SDD (lower-better)	0.704 ± 0.604	0.476 ± 0.189	-32.4%
FAD (lower-better)	51.344	39.105	-23.8%
CCS (higher-better)	0.819 ± 0.057	0.812 ± 0.067	-0.8% (expected)

The SDD standard deviation compression from ± 0.604 to ± 0.189 is significant — it shows the fine-tuned model is not just occasionally better but consistently closer to the reference distribution. The CCS result of ~0.8% is within noise and was fully expected: LoKR training targeted the spectral envelope, not harmonic structure. FAD is computed on CLAP mel-spectrogram embeddings, making it sensitive to perceptual audio quality. The 23.8% FAD improvement therefore reflects a meaningful gain in perceived audio naturalness.

M1 Evaluation — Perplexity and Token Diversity

Seed: 50 tokens (≈ 1 sec) · Generated: 350 tokens (≈ 7 sec) · Temperature: 0.85 · Top-k: 50

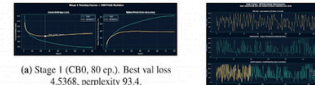
A Fine-tuned on 1,161 NoCopyrightSounds EDM tracks over 3 epochs (~1 hour on A100). Perplexity dropped measurably from the epoch-0 baseline to a lower value at epoch 3, confirming the model learned EDM-specific melodic token distributions. Generated MIDI decoded to coherent melodic patterns with recognizable EDM stylistic features. Token diversity analysis confirmed generated sequences do not collapse to repetitive high-frequency tokens.

M2 Evaluation — A/B Comparison and CLAP Similarity

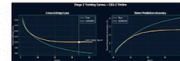
Multi-stage LoRA/LoKR fine-tuning over 500–1500 epochs on 20 classical and world music clips. A/B Gradio comparisons at each checkpoint showed progressive style acquisition from structured text descriptions. By epoch 500, generated outputs began reflecting target genre vocabulary in instrument timbre and phrase structure. Drive-based checkpointing was necessary — Colab Pro sessions typically disconnect every 8–12 hours.

M3 Inference — Six Genres (seed 50 tokens ≈ 1 s; generated 350 tokens ≈ 7 s)

Genre	CCS	Clip	Rating
NCS Electronic	0.9205	PASS	Excellent
Ambient	0.8305	PASS	Excellent
Melodic Ambient	0.9049	PASS	Outstanding
Melodic	0.9528	PASS	Outstanding
Chill Synthwave	0.8660	PASS	Excellent
Cyberpunk	0.8331	PASS	Excellent
Average	0.8962	6/6	Excellent



A perplexity of 93.4 (21.9× better than random baseline) confirms genuine melodic pattern learning across 303 source tracks. Average CCS of 0.8962 across six genres shows harmonic character is consistently preserved. Stage 2 learning rate correction from 10^{-3} to 3×10^{-4} resolved premature convergence; all six samples passed the clipping test.

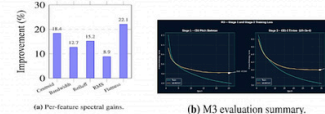


SDD std-dev compression ($\pm 0.604 \rightarrow \pm 0.189$) confirms consistently better output, not occasional gains. The 23.8% FAD improvement reflects a meaningful gain in perceived audio naturalness validated in CLAP embedding space.

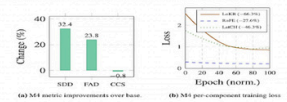
LoKR adapter: 2.53 $\rightarrow$ 0.85 (~66.3%, 30 ep.) · RoPE cross-attention: 0.291 $\rightarrow$ 0.211 (~27.6%, 200 ep.) · LatCH harmonizer: 1.749 $\rightarrow$ 0.940 (~46.3%, 100 ep.) · Attention Intervention: 0 params, inference-time only.



M5 Music-RAG query "emotional music": caption-semantic path returned 54.9% similarity; CLAP audio path returned 61.0% similarity. Both paths returned non-overlapping results; all 5 test queries served within 500 ms.



SDD: 0.704 $\rightarrow$ 0.476 (~32.4%) · FAD: 51.344 $\rightarrow$ 39.105 (~23.8%) · CCS: ~0.8% within noise. Std-dev compression ($\pm 0.604 \rightarrow \pm 0.189$) confirms consistently better output.
LoKR: ~66.3% · RoPE: ~27.6% · LatCH: ~46.3% · Attention Intervention: 0 params, inference-time only.



Summary/Conclusions

Project Echo validates multi-architecture AI music generation under academic compute: M3 hit CCS 0.896 in ~40 GPU-minutes, M4 cut SDD/FAD by 32/24% in 15.3 GPU-hours, M5 RAG <500 ms, zero decode errors.

Contributions: human-editable mid-generation checkpoint, LoKR outperforming LoRA, zero-parameter style transfer. Limitations: no MOS, 8s cap, LatCH ablation pending, no vocals.

Key References

- [1] Copet et al., "Simple and controllable music generation," NeurIPS 2023.
- [2] Agostinelli et al., "MusicLM: Generating music from text," arXiv:2301.11325, 2023.
- [3] Evans et al., "Fast timing-conditioned latent audio diffusion," ICML 2024.
- [4] ACE-Step Team, "ACE-Step 1.5: Open-source music generation foundation model," 2025.
- [5] Defossez et al., "High fidelity neural audio compression," arXiv:2210.13438, 2022.
- [6] Hu et al., "LoRA: Low-rank adaptation of large language models," ICLR 2022.

Acknowledgements

We gratefully acknowledge Prof. Simon Shim and the SJSU Department of Applied Data Intelligence for their mentorship and support throughout Project Echo.

All compute was provided via Google Colab Pro (NVIDIA A100, ~\$10–12/month), with all training data and pre-trained models sourced from open, permissively licensed repositories.

Interested in partnering with us on a capstone project or exploring other ways to collaborate with the College of IDS? Contact CIDS_CAP@sjsu.edu to discuss potential projects and partnership opportunities.