Neural Codec Language Model for Controllable Timbre Transfer in Music Synthesis

Sheldon Liu SHILONG@AMAZON.COM
Tianyu Liu XYZLIU@AMAZON.COM
Deepak Dalakoti DALAKOTI@AMAZON.COM
Adithya Suresh ADXTHYA@AMAZON.COM
Yueying Teng YYTENG@AMAZON.COM
Xuefeng Liu LIUXUEFE@AMAZON.COM
Atanu Roy ATANUROY@AMAZON.COM
Amazon Web Services Australia Pty. Ltd., 2 Park Street, NSW, Australia

Randeep Bhatia RANDEEP@SPLASHMUSIC.COM
Daniel Hatadi DANIELH@SPLASHMUSIC.COM
Prabhjeet Ghuman PRABH@SPLASHMUSIC.COM
Splash, Brisbane, QLD, Australia

Abstract

Neural codec language models have revolutionized speech synthesis but face significant challenges when adapted to music generation, particularly in achieving precise timbre control while preserving melodic content. We introduce **N**eural **C**ode **L**anguage **M**odel for **C**ontrollable **T**imbre **T**ransfer (**NCLMCTT**), a novel architecture that enables zero-shot instrument cloning through direct audio conditioning without explicit timbre learning. Our approach combines a 385M-parameter transformer for coarse musical structure modeling with a specialized upsampler for fine timbral detail, achieving flexible control through 1-5 second reference audio segments. We establish the first comprehensive benchmark dataset for controllable timbre transfer evaluation, comprising 62,500 high-fidelity samples across 50 synthesizer presets with ground truth targets. Extensive experiments demonstrate substantial improvements over the TokenSynth baseline: 27.1% reduction in SI-SDR, 50.9% in Mel Distance, and 59.4% in STFT Distance, while maintaining strong melodic coherence (Chroma Similarity: 0.85). Our method achieves robust zero-shot generalization, with performance on unseen instrument presets matching that of seen presets. Ablation studies confirm that extended reference audio duration (40.8% improvement), cross-attention mechanisms (11.9% improvement), and increased model capacity contribute meaningfully to overall performance. By separating melodic content from timbral characteristics and enabling implicit timbre control, NCLMCTT provides both immediate practical value for music creators and a methodological foundation for advancing controllable neural audio synthesis.

Keywords: Neural codec language models, Timbre transfer, Controllable music synthesis, Zero-shot generalization, Audio generation

1. Introduction

The democratization of music creation through AI has reached a critical juncture. While commercial systems like Suno and Udio generate impressive musical compositions Nugroho and Manggala (2024), the field lacks precise controllable synthesis mechanisms that separate

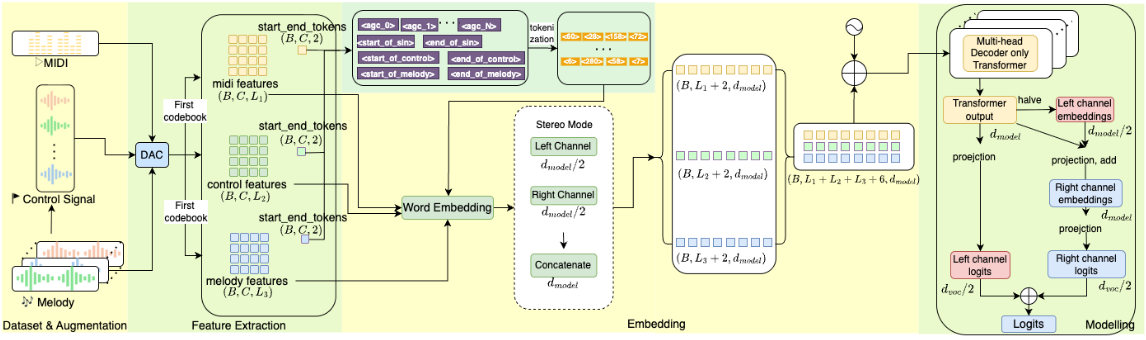


Figure 1: NCLMCTT architecture for controllable timbre transfer. The pipeline processes MIDI input and control signals through four stages: (1) Dataset augmentation and DAC conversion, (2) Feature extraction from MIDI, control (reference audio), and melody (source audio) into tokenized representations, (3) Embedding with positional encoding, and (4) Transformer-based modeling with multi-head self-attention and cross-attention mechanisms.

melodic content from timbral characteristics. Current text-to-music models struggle with fine-grained instrument control due to natural language ambiguity Schneider et al. (2024); Agostinelli et al. (2023).

Neural codec language models revolutionized speech synthesis by treating audio generation as discrete token prediction Wang et al. (2023). Recent efforts to adapt these models to music generation show promise but face critical challenges: existing approaches either rely on token-level manipulation without explicit conditioning mechanisms or require pre-trained timbre encoders that limit flexibility. The need for music-specific architectures that can leverage the efficiency of neural codec models while providing precise timbre control remains largely unaddressed.

The evaluation crisis compounds these challenges. Current metrics like Fréchet Audio Distance (FAD) show poor correlation with human judgment, while commercial systems now outperform reference datasets Grötschla et al. (2025). The lack of standardized protocols for controllable synthesis has hindered rigorous comparison and slowed progress in the field.

Our contributions advance neural codec language models for music synthesis across technical and methodological dimensions. We introduce **N**eural **C**ode **L**anguage **M**odel for **C**ontrollable **T**imbre **T**ransfer (**NCLMCTT**)(Fig. 1), featuring flexible control signal durations, zero-shot instrument control through direct audio conditioning. We establish the first comprehensive benchmark dataset for controllable timbre transfer evaluation, comprising 62,500 high-fidelity samples with standardized metrics. Our empirical validation demonstrates substantial improvements over the TokenSynth Kim et al. (2025) baseline: 27.1% reduction in Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), 50.9% in Mel Distance, and 59.4% in Short-Time Fourier Transform (STFT) Distance, while maintaining strong melodic preservation. These improvements are achieved through music-specific architectural modifications including hierarchical token generation, explicit cross-attention mechanisms for timbre conditioning, and flexible reference audio duration from 1 to 5 seconds.

We provide open-source access to our curated dataset, evaluation protocols, and complete implementation, establishing a reproducible foundation for advancing controllable neural music synthesis research. By positioning neural codec language models as a bridge between symbolic music representation and audio synthesis, our work provides both immediate practical value for music creators and a methodological foundation for advancing controllable audio generation research.

2. Related Work

Neural Codec Language Models: VALL-E Wang et al. (2023) pioneered neural codec language models for speech synthesis, using EnCodec's hierarchical representations for zero-shot voice cloning with 3-second enrollment. VALL-E 2 Chen et al. (2024) achieved human parity through repetition-aware sampling and grouped code modeling, reducing word error rates by 50%. AudioLM Borsos et al. (2023) introduced semantic-acoustic decomposition using w2v-BERT and SoundStream tokens, enabling controllable synthesis without text transcripts but requiring computationally expensive cascaded models. **Text-to-Music Generation**: MusicLM Agostinelli et al. (2023) adapts AudioLM's hierarchical approach but struggles with precise instrument control due to language ambiguity. MusicGen Copet et al. (2023) revolutionized efficiency through single-stage architecture with token interleaving across EnCodec's 4 codebooks, achieving superior performance while reducing computational requirements. Moûsai Schneider et al. (2024) employs latent diffusion with 64× compression for high-quality stereo generation but requires complex sampling procedures. **Controllable Music Synthesis and Audio Codecs**: NSynth Engel et al. (2017) established timbre control through 16-dimensional embeddings, while DDSP approaches Engel et al. (2020) enable interpretable control through synthesizer parameters. However, most methods require explicit timbre learning and lack zero-shot generalization. Recent audio codecs have achieved extreme compression: EnCodec Défossez et al. (2022) provides 16-32× compression, Descript Audio Codec (DAC) Kumar et al. (2023) achieves 90× ratios with superior quality, and WavTokenizer reduces audio to 40-75 tokens per second Ji et al. (2024). **Instrument Cloning**: TokenSynth Kim et al. (2025) performs zero-shot polyphonic instrument cloning using CLAP-conditioned transformers that generate DAC tokens autoregressively, enabling text-guided timbre manipulation through cross-modal embedding interpolation. However, the method relies on pretrained CLAP embeddings for timbre conditioning, requiring explicit timbre representations learned during pretraining. In contrast, our approach achieves adaptive timbre transfer without explicit timbre learning, enabling more flexible generalization. We benchmark against TokenSynth as the most closely related work to our framework. **Evaluation and Our Approach**: Current evaluation methodologies suffer from significant limitations, with FAD showing poor correlation with human judgment and lack of standardized protocols for controllable synthesis Gui et al. (2024); Grötschla et al. (2025). Our work addresses these gaps through music-specific architectural components and audio-based conditioning for zero-shot generalization. Most importantly, we introduce the first comprehensive benchmark specifically designed for controllable timbre transfer, enabling systematic evaluation of both melodic preservation and timbral fidelity.

3. Problem Formulation

Task Definition: We address timbre-conditioned melody synthesis in neural audio generation, where timbre control is derived directly from reference audio rather than text descriptions or predefined instrument categories. Given a dataset containing 1,250 MIDI melodies $\mathcal{M} = \{m_1, \ldots, m_{1250}\}$, 50 synthesizer presets $\mathcal{P} = \{p_1, \ldots, p_{50}\}$, and corresponding rendered waveforms $\mathcal{W} = \{w_{i,j} | \ i \in [1, 1250], j \in [1, 50]\}$, we formalize the task as learning a mapping function f_θ that transforms a MIDI melody m_i and a reference audio snippet $c_{a,j}$ (1-5s crop from $w_{a,j}$) into a synthesized waveform:

$$\hat{w}_{i,j} = f_\theta(m_i, c_{a,j}) \tag{1}$$

Timbre Transfer Mechanism: The reference audio $c_{a,j}$ serves as the sole source of timbre information, enabling the model to extract and transfer timbral characteristics without relying on intermediate representations such as text embeddings or explicit instrument labels. When $a \neq i$, the control signal contains different melodic content compared to the input MIDI m_i, enabling adaptive timbre transfer without explicit timbre learning.

Evaluation Framework: Critically, this formulation enables objective evaluation by providing ground truth waveforms $w_{i,j}$ for direct comparison with generated outputs $\hat{w}_{i,j}$ across multiple metrics (see Section 5 for details), eliminating the need for costly and potentially inconsistent human subjective evaluations that plague many timbre transfer and music generation benchmarks. We measure success through four complementary metrics: **SI-SDR** (↓) for time-domain fidelity, **MEL Distance** (↓) for perceptual quality, **STFT Distance** (↓) for time-frequency accuracy, and **Chroma Similarity** (↑) for melodic preservation, ensuring both timbral fidelity and melodic accuracy are rigorously assessed.

4. Proposed Method

4.1. Preprocessing and Feature Extraction

Our pipeline transforms 1,250 MIDI melodies and 50 synthesizer presets into 62,500 unique audio waveforms. Training triplets consist of $(mi_x, \mathbf{C}, me_x)$ where mi_x is input MIDI melody, $\mathbf{C}$ is tiled control signal from potentially any melody using preset x, and me_x is target waveform with preset consistency. DAC Encoder processes three modalities (MIDI files, complete audio, and cropped audio) into features with shape (B, L, C), where B is batch size, L is number of codebooks (9 when using DAC), and C is sequence length.

4.2. Architecture Overview

Our NCLMCTT architecture separates coarse musical structure generation from fine timbral detail synthesis. The first stage employs a 385M-parameter transformer-based LLM for autoregressive coarse codebook token generation, establishing musical structure while incorporating timbre conditioning. The second stage utilizes a specialized upsampling module for non-autoregressive fine token prediction, transforming coarse structure into high-fidelity audio.

Stage 1 - LLM for Coarse Token Generation: Our LLM employs a decoder-only transformer ($L = 24$ layers, $d_{model} = 1024$, $h = 16$ heads) designed for musical token prediction. The model autoregressively generates coarse codebook sequences:

$$p(z^1_{1:T}) = \prod_{t=1}^{T} p(z^1_t | z^1_{<t}, m_i, c_{a,j}) \tag{2}$$

For stereo audio, we implement channel dependency modeling. Input tokens are processed through learnable embeddings with sinusoidal positional encodings. Our model incorporates control tokens enabling flexible conditioning, with cross-attention layers integrating timbre control signals with MIDI input. Training employs standard autoregressive language modeling with cross-entropy loss, while inference uses temperature scaling and top-k filtering with Gumbel-max sampling.

Stage 2 - Specialized Upsampling Module: The upsampling module transforms coarse tokens into high-fidelity audio through non-autoregressive fine token prediction. Given coarse tokens $z^1_{1:T}$, the upsampler predicts fine tokens through:

$$P_\theta(\hat{z}^2_{1:T}, \ldots, \hat{z}^9_{1:T} | z^1_{1:T}) = \prod_{i=2}^{9} P_\theta(\hat{z}^i_{1:T} | z^1_{1:T}, \hat{z}^{2:i-1}_{1:T}) \tag{3}$$

The module incorporates three conditioning types: **Metrical Conditioning** using beat phase information, **Harmonic Conditioning** through pitch class histograms and root note embeddings, and **Channel Conditioning** for stereo generation. Training employs masked token prediction with selective masking of fine tokens while preserving coarse structure.

4.3. Data Augmentation and Training

We implemented a data augmentation strategy (Figure 2) creating triplets of MIDI input mi_x, tiled control signal $\mathbf{C}$, and target melody me_x. Control signals are extracted crops (t_c seconds) from target waveforms using the same synthesizer preset but potentially different melodies, then tiled to match the target length. This technique decouples timbre from specific melodic content. By pairing each MIDI sequence with different preset renderings, we expanded our dataset from 62,500 to approximately 1.25 million samples (strategically selected from a theoretical 3 million possibilities), while varying crop lengths (1-5 seconds) to teach the model to extract timbral characteristics from control signals of different duration.

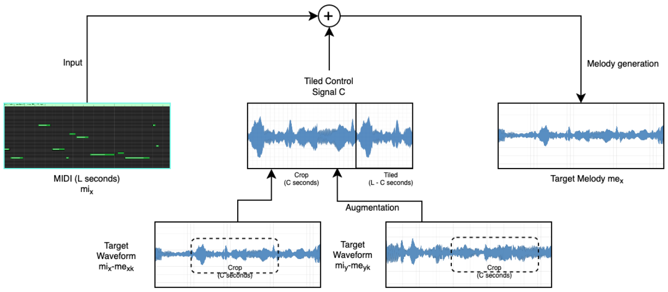


Figure 2: Data augmentation strategy for NCLMCTT

5. Experiments

5.1. Training Configuration

5.1.1. First Codebook Model Training

We trained our transformer-based model for first codebook token prediction using distributed data parallel across 8 NVIDIA L40S GPUs, completing in 4-6 hours with a per-GPU batch size of 16-24 (global batch size 128-192) for 200-500 steps per epoch. The training used a learning rate of 1×10^{-4} with 200 warm-up steps in cosine scheduling, implemented bfloat16 mixed-precision via PyTorch AMP for efficient memory usage, and employed Adam optimizer with CrossEntropyLoss, while monitoring all metrics through TensorBoard to ensure training stability.

5.1.2. Upsampler Training

The upsampler predicts fine codebook tokens (z_2 through z_9) conditioned on the first codebook z_1 using masked token prediction. We selectively mask tokens in codebooks z_2-z_9 while preserving z_1, with masking rates from 0.5-99.3%. The model optimizes cross-entropy loss per codebook level via non-autoregressive parallel prediction. Implementation uses a 20-layer transformer (dimension 1280) with AdamW optimizer (learning rate 3e-5), trained on 2048-token segments (∼24s) using bfloat16 precision and an 80/20 train/test split (350 test samples).

5.2. Identifying the First Codebook as Bottleneck

To validate our hierarchical design, we evaluated upsampler performance when conditioned on ground truth (GT) first codebook tokens. Figure 3 shows results across all 50 presets. Upsampling from GT tokens achieves near-perfect reconstruction (median Chroma Similarity: 0.9879, SI-SDR: -2.7 dB, Mel Distance: 0.9234, STFT Distance: 1.711), demonstrating that the upsampler effectively generates fine timbral details when provided with accurate coarse structure. This validates our architectural decomposition: the first codebook captures the primary bottleneck for timbre transfer quality. Based on this analysis, we use end-to-end evaluation for baseline comparison (Section 5.5.1) and first codebook evaluation for architectural analysis (Section 6).

5.3. Evaluation Metrics

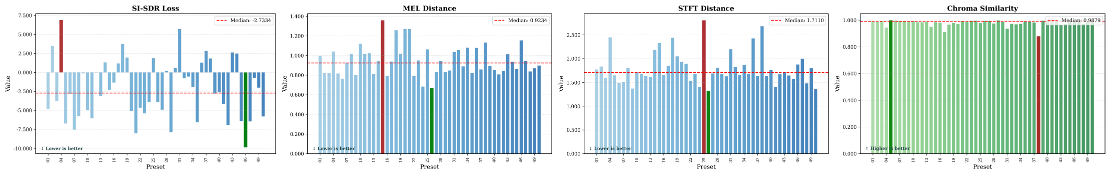
To assess our method quantitatively, we employ four complementary metrics capturing different aspects of audio quality:

SI-SDR(↓): Scale-Invariant Signal-to-Distortion Ratio measures time-domain fidelity while being invariant to scaling (the SI-SDR is negated, so lower value indicates better performance):

$$\text{SI-SDR}(x, \hat{x}) = -10 \log_{10} \frac{|\alpha x|^2}{|\alpha x - \hat{x}|^2}, \text{ where } \alpha = \frac{\hat{x}^T x}{|x|^2} \tag{4}$$

MEL Distance(↓): Evaluates perceptual differences in the mel-frequency domain using multi-scale mel spectrograms:

$$d_{\text{MEL}}(x, \hat{x}) = \frac{1}{TM} \sum_{t,m} |\text{MEL}_{t,m}(x) - \text{MEL}_{t,m}(\hat{x})|_1 \tag{5}$$



(a) (b) (c) (d)

Figure 3: Upsampling from ground-truth first codebook tokens across 50 presets. Near-perfect reconstruction (median Chroma Similarity: 0.9879) confirms first codebook modeling as the key bottleneck in timbre transfer. (a) SI-SDR, (b) MEL Distance, (c) STFT Distance, (d) Chroma Similarity.

STFT Distance(↓): Measures time-frequency representation discrepancies:

$$d_{\text{STFT}}(x, \hat{x}) = \frac{1}{TF} \sum_{t,f} |\text{STFT}_{t,f}(x) - \text{STFT}_{t,f}(\hat{x})|_1 \tag{6}$$

Chroma Similarity(↑): Quantifies musical similarity through chromagram cosine similarity:

$$\text{ChromaCosSim}(X, \hat{X}) = \frac{1}{T} \sum_{t} \frac{x_t \cdot \hat{x}_t}{|x_t|_2 \cdot |\hat{x}_t|_2 + \epsilon} \tag{7}$$

We implemented SI-SDR, MEL Distance, and STFT Distance using the `audiotools` library[1], specifically from the `spectral.py` and `distance.py` modules. For SI-SDR, we used the default parameters with zero-mean normalization and mean reduction across batches. For MEL Distance, we employed a multi-scale approach using two resolutions (150 and 80 mel bands) with window lengths of 2048 and 512, combining both magnitude and log-magnitude L1 losses. Similarly, the STFT Distance was calculated using multi-scale STFT with window lengths of 2048 and 512, also combining magnitude and log-magnitude losses with equal weighting.

5.4. Baseline Selection

Table 1 compares existing controllable music synthesis methods across key functional dimensions to identify appropriate baselines for benchmarking. **Excluded Methods**: Speech synthesis methods (VALL-E Wang et al. (2023), VALL-E 2 Chen et al. (2024), AudioLM Borsos et al. (2023)) are optimized for linguistic structure rather than musical characteristics, operating on phoneme-aligned inputs and shorter temporal contexts suitable for utterances but insufficient for musical phrases. Text-to-music methods (MusicLM Agostinelli et al. (2023), MusicGen Copet et al. (2023), Moûsai Schneider et al. (2024)) cannot perform reference-based timbre transfer as they rely solely on text descriptions, which introduce ambiguity for precise instrument specification. **Selected Baseline**: We benchmark against *TokenSynth* Kim et al. (2025), which represents the state-of-the-art in zero-shot instrument synthesis through discrete token manipulation. TokenSynth is the only existing method that

[1] https://github.com/descriptinc/audiotools

Table 1: Functional comparison of controllable music synthesis methods. NCLMCTT uniquely combines reference audio conditioning, flexible control length, and zero-shot transfer without explicit timbre learning.

Method	Control Signal Type	No Explicit Timbre Learning	Flexible Control Length	Zero-shot Transfer
VALL-E Wang et al. (2023)	Reference Audio	✓	✗	✓
VALL-E 2 Chen et al. (2024)	Reference Audio	✓	✗	✓
AudioLM Borsos et al. (2023)	Reference Audio	✗	✗	✗
MusicLM Agostinelli et al. (2023)	Text	✓	✓	✗
MusicGen Copet et al. (2023)	Text	✓	✓	✗
Moûsai Schneider et al. (2024)	Text	✓	✓	✗
TokenSynth Kim et al. (2025)	Text + Reference	✓	✓	✓
NCLMCTT (Ours)	**Reference Audio**	✓	✓	✓

combines reference audio control with zero-shot transfer capabilities and flexible control length, making it functionally most similar to our approach and enabling direct comparison of timbre transfer quality.

Architectural Inspiration: While VALL-E Wang et al. (2023) inspired our hierarchical coarse-to-fine generation strategy, NCLMCTT introduces substantial music-specific modifications: (1) cross-attention mechanisms for explicit timbre conditioning beyond concatenation -based approaches; (2) flexible reference duration (1-5s) for varying musical phrase lengths; (3) MIDI-based melodic features and extended temporal contexts optimized for musical structure. These deviations address fundamental differences between speech synthesis (phoneme-to-audio with speaker identity) and music synthesis (MIDI-to-audio with timbral control), making TokenSynth the appropriate benchmark for timbre transfer evaluation rather than adapting speech models.

5.5. Benchmarking

5.5.1. Comparison with TokenSynth

Table 2 presents the quantitative comparison between TokenSynth and NCLMCTT. Our method achieves substantial improvements in timbral fidelity across all spectral metrics: 27.1% lower SI-SDR, 50.9% lower Mel Distance, and 59.4% lower STFT Distance. These improvements demonstrate that our direct audio conditioning approach enables significantly more accurate timbre replication than TokenSynth's CLAP embedding-based method.

TokenSynth achieves marginally higher Chroma Similarity (0.878 vs. 0.850, a 3.2% difference), indicating slightly better melodic preservation. However, both methods maintain strong melodic coherence above 0.85, suggesting that NCLMCTT successfully balances timbre transfer with melodic integrity. The results validate our hypothesis that avoiding pretrained timbre encoders enables more flexible and accurate timbre transfer, achieving 50-59% improvements in spectral distance metrics while maintaining competitive melodic performance.

5.6. Zero-shot Generalization Performance

NCLMCTT demonstrates robust generalization to unseen instrument presets, validating its ability to transfer learned timbre modeling capabilities to novel sounds without additional training. Figure 4 presents a comprehensive comparison between zero-shot performance on

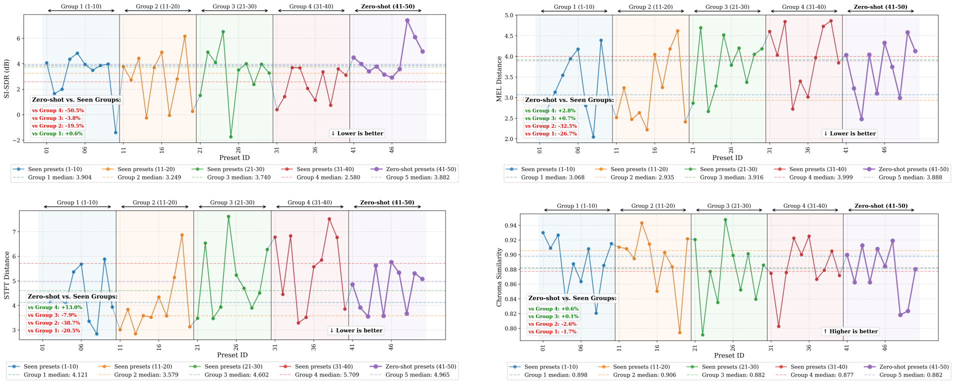


Figure 4: Zero-shot performance comparison between seen training presets (groups 1-10, 11-20, 21-30, 31-40) and unseen test presets (41-50). Box plots show distribution of SI-SDR, Mel Distance, STFT Distance, and Chroma Similarity across preset groups, demonstrating robust generalization to novel instrument timbres.

unseen presets (presets 41-50) and performance on four groups of seen presets from the training set (presets 1-10, 11-20, 21-30, 31-40).

The results reveal strong zero-shot generalization across all metrics. For SI-SDR and Mel Distance, unseen presets achieve performance comparable to seen presets, with distributions largely overlapping. STFT Distance shows particularly robust generalization, with zero-shot performance matching or exceeding several seen preset groups. Chroma Similarity maintains consistent performance across both seen and unseen presets, indicating that melodic preservation is not degraded when transferring to novel timbres.

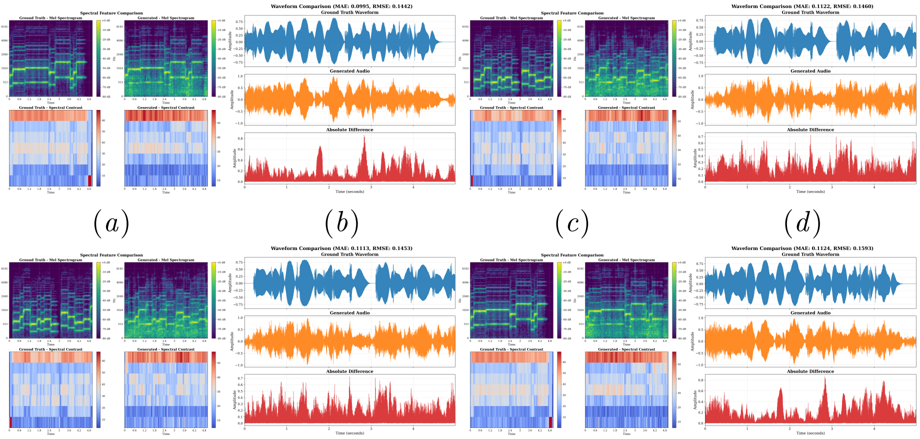
5.7. Qualitative Analysis

Figure 5 presents detailed spectral and waveform visualizations for four representative samples, demonstrating NCLMCTT's accurate reconstruction capabilities across diverse timbral characteristics. Spectral comparisons (panels a, c, e, g) reveal close alignment between generated and ground truth spectrograms, with predicted outputs preserving harmonic structure, formant characteristics, and temporal spectral evolution. Chromagram analysis (bottom rows of spectral panels) confirms strong melodic preservation across all samples, with Chroma Similarity scores ranging from 0.96 to 0.99. Waveform comparisons (panels

Table 2: Comparison of audio quality metrics between TokenSynth and NCLMCTT. Lower values indicate better performance for SI-SDR, Mel Distance, and STFT Distance, while higher values are better for Chroma Similarity. Bold indicates best performance.

Metric	**TokenSynth**	**NCLMCTT (Ours)**	**Improvement**
SI-SDR ↓	29.22	**21.30**	27.1%
Mel Distance ↓	3.69	**1.81**	50.9%
STFT Distance ↓	6.41	**2.60**	59.4%
Chroma Similarity ↑	**0.878**	0.850	-3.2%

b, d, f, h) show faithful reproduction of temporal envelopes, though absolute difference plots reveal notable deviations concentrated at note onsets where precise transient modeling remains challenging—a common limitation in autoregressive token-based generation. Additionally, subtle high-frequency artifacts are visible in some spectrograms (particularly Sample 1), suggesting that extremely fine timbral details occasionally suffer minor degradation. Despite these localized imperfections, the consistent overall performance across samples (SI-SDR: -3.64 to -4.72 dB, Mel Distance: 2.35-2.49, STFT Distance: 3.05-3.26) validates NCLMCTT's robust generalization, successfully transferring timbre while maintaining melodic integrity across varying melodic patterns and instrumental characteristics without major artifacts such as spectral smearing or temporal jitter.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5: Audio visualization for all samples. Each sample shows spectral features (left) and waveform (right). Sample 1 (SI-SDR: -4.72, MEL: 2.35, STFT: 3.05, Chroma: 0.98), Sample 2 (SI-SDR: -3.69, MEL: 2.48, STFT: 3.19, Chroma: 0.96), Sample 3 (SI-SDR: -3.68, MEL: 2.48, STFT: 3.19, Chroma: 0.98), Sample 4 (SI-SDR: -3.64, MEL: 2.49, STFT: 3.26, Chroma: 0.99)

6. Ablation Studies

To validate our design choices, we conduct ablation studies examining control signal duration, cross-attention mechanisms, and model capacity. Table 3 presents first codebook reconstruction performance across different NCLMCTT configurations. **Control Signal Duration.** We evaluate reference audio lengths from 1s to 5s with fixed architecture (0.3B parameters, no cross-attention). Performance improves consistently with longer references: nclmctt_5s achieves 40.8% lower SI-SDR than nclmctt_1s (5.27 → 3.12 dB), with corresponding improvements in Mel Distance (3.81 → 3.52) and Chroma Similarity (0.86 → 0.88). This validates that extended reference signals provide richer timbral information for more accurate transfer. **Cross-Attention Mechanism.** Adding explicit cross-attention layers (nclmctt_cross_attn) further reduces SI-SDR from 3.12 to 2.75 dB (11.9% improvement over nclmctt_5s), with gains in Mel Distance (3.52 → 3.46). This demonstrates that explicit attention to reference audio enhances timbre conditioning beyond

Table 3: Ablation study of NCLMCTT design choices on first codebook reconstruction. Configurations vary by control signal duration (1s-5s), cross-attention mechanism, and model size (0.3B vs. 1.2B parameters). Best values are highlighted in bold.

Model	**SI-SDR (dB) ↓**	**Mel Dist. ↓**	**STFT Dist. ↓**	**Chroma Sim. ↑**
Control Signal Duration (0.3B params, no cross-attn)				
nclmctt_1s	5.27 ± 4.18	3.81 ± 1.30	4.70 ± 1.74	0.86 ± 0.08
nclmctt_2s	4.50 ± 2.92	3.77 ± 1.13	4.72 ± 1.65	0.86 ± 0.08
nclmctt_3s	4.37 ± 2.66	3.53 ± 0.95	4.51 ± 1.39	0.87 ± 0.06
nclmctt_4s	4.33 ± 3.57	3.70 ± 1.19	4.65 ± 1.63	0.87 ± 0.07
nclmctt_5s	3.12 ± 2.51	3.52 ± 1.03	4.54 ± 1.55	0.88 ± 0.06
Architecture Enhancements (5s reference)				
nclmctt_cross_attn	2.75 ± 2.63	3.46 ± 1.02	4.55 ± 1.56	0.88 ± 0.06
nclmctt_cross_attn_large	**2.75 ± 2.13**	**3.28 ± 0.87**	**4.41 ± 1.32**	**0.89 ± 0.06**

concatenation-based approaches. **Model Capacity.** Scaling from 0.3B to 1.2B parameters (nclmctt_cross_attn_large) maintains SI-SDR (2.75 dB) while achieving best overall performance: Mel Distance improves to 3.28 (5.2% better), STFT Distance to 4.41 (3.1% better), and Chroma Similarity to 0.89 (1.1% better). Increased capacity enables better spectral modeling and melodic preservation without overfitting.

7. Conclusion

We introduced NCLMCTT, a neural codec language model that advances controllable timbre transfer through implicit audio conditioning without pretrained timbre encoders and the first comprehensive benchmark dataset for systematic evaluation. Compared to TokenSynth, NCLMCTT achieves substantial improvements in timbral fidelity (27.1% reduction in SI-SDR, 50.9% in Mel Distance, 59.4% in STFT Distance) while maintaining strong melodic coherence (Chroma Similarity: 0.85). Zero-shot evaluation on unseen presets confirms robust generalization without performance degradation. Ablation studies reveal that control signal duration provides the largest impact (40.8% improvement), followed by cross-attention mechanisms (11.9%) and model scaling (up to 5.2%). The analysis demonstrates that first codebook modeling represents the primary bottleneck, with upsampling from ground truth tokens achieving near-perfect reconstruction. By establishing rigorous evaluation protocols and providing open-source access to our dataset and implementation, NCLMCTT serves as both a practical tool for music creators and a methodological foundation for advancing controllable neural audio synthesis.

8. Limitations and Future Work

While NCLMCTT demonstrates strong performance, several directions warrant further investigation: incorporating explicit pitch-aware mechanisms to close the 3.2% melodic preservation gap with TokenSynth and scaling to longer musical phrases with real-time generation capabilities. By establishing a rigorous evaluation framework and providing open-source access to our dataset, protocols, and implementation, we aim to accelerate progress in controllable neural music synthesis. Our work demonstrates that neural codec language models can effectively bridge symbolic music representation and audio synthesis, enabling precise timbre control while maintaining melodic integrity, positioning NCLMCTT as both a practical tool for music creators and a methodological foundation for advancing controllable audio generation research.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*, 2024.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720, 2023.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International conference on machine learning*, pages 1068–1077. PMLR, 2017.

Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020.

Florian Grötschla, Ahmet Solak, Luca A Lanzendörfer, and Roger Wattenhofer. Benchmarking music generation models and metrics via human preference studies. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1331–1335. IEEE, 2024.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024.

Kyungsu Kim, Junghyun Koo, Sungho Lee, Haesun Joung, and Kyogu Lee. Tokensynth: A token-based neural synthesizer for instrument cloning and text-to-instrument. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993, 2023.

Y Yogi Tegar Nugroho and P Paulus Metta Dwi Manggala. The use of ai in creating music compositions: A case study on suno application. In *7th Celt International Conference (CIC 2024)*, pages 177–189. Atlantis Press, 2024.

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8050–8068, 2024.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.