Adapting SoundStorm for Environmental Sound Generation

Kathryn Chen, Zhanliang Wang, Tripti Tripathi
CIS 7000 Final Project
University of Pennsylvania
{fdshfg, aaronwzl, triptit}@upenn.edu

December 11, 2025

Abstract

We investigate the adaptation of SoundStorm, a state-of-the-art parallel audio generation model originally designed for speech synthesis, to environmental sound generation. Through comprehensive experiments on ESC-50, we evaluate both the generative and discriminative capabilities of the fine-tuned model. Our quantitative analysis using eight acoustic metrics reveals systematic failure modes: high-frequency attenuation with 67% reduction in zero-crossing rate, temporal over-smoothing evidenced by 55% increase in signal-to-noise ratio, and diminished spectral diversity reflected in 9.5% lower spectral entropy. Classification experiments demonstrate that while SoundStorm representations can be adapted for discriminative tasks, achieving 79.6% accuracy on ESC-50, they underperform task-specific models by approximately 9 percentage points. Despite quality limitations, our results validate that parallel masked refinement successfully transfers to non-speech domains, generating temporally coherent audio. We provide detailed analysis of the domain gap between speech and environmental sounds, identifying codec limitations, training biases, and architectural constraints as primary factors limiting generation quality. Our work establishes a rigorous evaluation framework for cross-domain audio generation and offers concrete directions for improving parallel synthesis of diverse acoustic phenomena.

Code: `https://github.com/ZhanliangAaronWang/CIS7000_Final_Project`

Contents

1 Introduction **3**

2 Related Work **4**
2.1 Neural Audio Generation 4
2.2 Masked Generative Modeling 4

2.3 Neural Audio Codecs 4
2.4 Environmental Sound Synthesis 4

3 Background: SoundStorm Architecture **5**
3.1 System Overview 5
3.2 Bidirectional Conformer Encoder 5
3.3 MaskGIT-Based Parallel Decoding 6
3.4 SoundStream Neural Codec 7
3.5 Decoder Pipeline 8

4 Experimental Setup **9**
4.1 Datasets 9
4.2 Model Configuration 9
4.3 Training Protocol 9
4.4 Evaluation Methodology 9

5 Results **10**
5.1 Classification Performance 10
5.2 Audio Generation Quality 11
5.2.1 Training Dynamics 11
5.2.2 Quantitative Acoustic Analysis 12
5.2.3 Statistical Significance 14
5.2.4 Pairwise Similarity Analysis 16
5.2.5 Qualitative Visual Analysis 16

6 Discussion **19**
6.1 Failure Mode Analysis 19
6.2 Domain Gap Characterization 20
6.3 Implications and Future Directions 21
6.4 Validation of Parallel Generation 21

7 Limitations **22**

8 Conclusion **22**

1 Introduction

Neural audio generation has witnessed remarkable progress with the development of transformer-based architectures and neural audio codecs. While autoregressive models have achieved impressive quality, their sequential generation process imposes significant computational constraints, particularly for long-form synthesis. SoundStorm [1] addresses this limitation through parallel decoding, achieving speech synthesis quality comparable to AudioLM while generating 30 seconds of audio in approximately 2 seconds on TPU-v4 hardware—a speedup exceeding 100× relative to real-time autoregressive generation.

The success of SoundStorm on speech synthesis raises a fundamental question regarding the generalizability of parallel generation frameworks: can models trained on speech successfully adapt to environmental sounds, which exhibit markedly different acoustic characteristics? Speech demonstrates quasi-periodic structure with identifiable formants, phonetic units, and spectral energy concentrated primarily between 300-3500 Hz. Environmental sounds, conversely, encompass stochastic textures like rain and wind, transient impulses such as door slams and glass breaking, and complex mechanical sounds with broadband spectral characteristics extending beyond 5 kHz. This acoustic diversity presents a substantially more challenging test of parallel generation capabilities.

Understanding domain transfer from speech to environmental audio carries both theoretical and practical significance. Theoretically, successful transfer would validate that the core components of SoundStorm—bidirectional Conformer encoders, MaskGIT-style masked prediction, and hierarchical residual vector quantization—capture domain-invariant principles of acoustic modeling. Systematic failures would illuminate architectural biases and domain-specific requirements. Practically, environmental sound generation enables applications in multimedia production, data augmentation for acoustic scene classification, accessibility technologies, and architectural acoustics simulation.

We investigate this domain transfer through comprehensive experiments on ESC-50, a standardized environmental sound dataset containing 2,000 five-second recordings across 50 semantic categories. Our study evaluates both generative quality through acoustic metrics and discriminative capability through classification accuracy. We identify specific failure modes, characterize the speech-to-environmental domain gap, and provide recommendations for improving cross-domain audio synthesis.

We provide three major contributions in this report. First, we provide the first systematic evaluation of SoundStorm on environmental sound generation, employing eight acoustic metrics, statistical testing, and visual analysis. Second, we demonstrate that SoundStorm representations can be adapted for audio classification, achieving 79.6% accuracy on ESC-50 while identifying a performance gap relative to task-specific architectures. Third, we characterize fundamental differences between speech and environmental audio that inform future architectural design for general-purpose audio generation systems.

2 Related Work

2.1 Neural Audio Generation

Autoregressive models have dominated neural audio generation, with WaveNet [2] pioneering the direct modeling of raw waveforms through dilated causal convolutions. Subsequent work introduced hierarchical generation schemes, with VQ-VAE [3] and Jukebox [4] operating on discrete latent representations. AudioLM [5] combined semantic and acoustic tokens from self-supervised models, achieving high-quality speech continuation and music generation through autoregressive modeling at multiple temporal scales.

While autoregressive approaches achieve impressive quality, their sequential nature imposes fundamental computational constraints. Recent work has explored parallel and iterative refinement methods to accelerate generation. Parallel WaveGAN [6] introduced adversarial training for non-autoregressive vocoding. DiffWave [7] and WaveGrad [8] applied denoising diffusion to waveform generation, trading sequential dependence for iterative refinement.

2.2 Masked Generative Modeling

Masked token prediction, originally developed for natural language processing with BERT [9], has been successfully adapted to other domains. MaskGIT [10] demonstrated that bidirectional transformers with iterative confidence-based unmasking could generate high-quality images in constant-time steps. This approach avoids the quadratic complexity of diffusion models while maintaining competitive quality. MaskGIT's success motivated its application to discrete audio tokens in SoundStorm.

2.3 Neural Audio Codecs

High-quality discrete audio representations are essential for token-based generation. SoundStream [11] introduced a neural codec combining convolutional autoencoders with residual vector quantization (RVQ), achieving perceptual quality superior to traditional codecs at comparable bitrates. Encodec [12] extended this architecture with improved training procedures and multi-resolution discriminators. These codecs provide the discrete token space on which SoundStorm operates.

2.4 Environmental Sound Synthesis

Environmental sound generation has received less attention than speech and music synthesis. Traditional approaches rely on concatenative synthesis or physical modeling. Recent neural approaches include WaveGAN [13] for general audio synthesis and GANSynth [14] for musical instrument sounds. However, most work focuses on discriminative tasks like environmental sound classification [15, 16] rather than generation. Our work provides the first systematic study of adapting a speech-trained parallel generation model to environmental sounds.

3 Background: SoundStorm Architecture

We provide a technical exposition of SoundStorm's architecture, focusing on components essential for understanding our experimental setup and results.

3.1 System Overview

SoundStorm consists of three primary components: a neural audio codec (SoundStream or Encodec) that converts waveforms to discrete RVQ tokens, a bidirectional Conformer encoder that models dependencies between tokens, and a parallel masked decoder that iteratively refines masked predictions. Given an audio waveform $\mathbf{y} \in \mathbb{R}^N$, the codec encoder produces a sequence of continuous embeddings which are quantized through Q residual quantizers, yielding discrete tokens $\mathbf{x} \in \{1, \ldots, K\}^{T \times Q}$ where T is the number of frames, K is the codebook size, and Q is the number of quantizer levels. The Conformer processes these tokens through bidirectional self-attention and convolution, producing contextualized representations. During generation, the decoder iteratively unmasks high-confidence token predictions until the complete sequence is reconstructed.

Figure 1 illustrates the high-level architecture of SoundStorm as presented in the original paper.

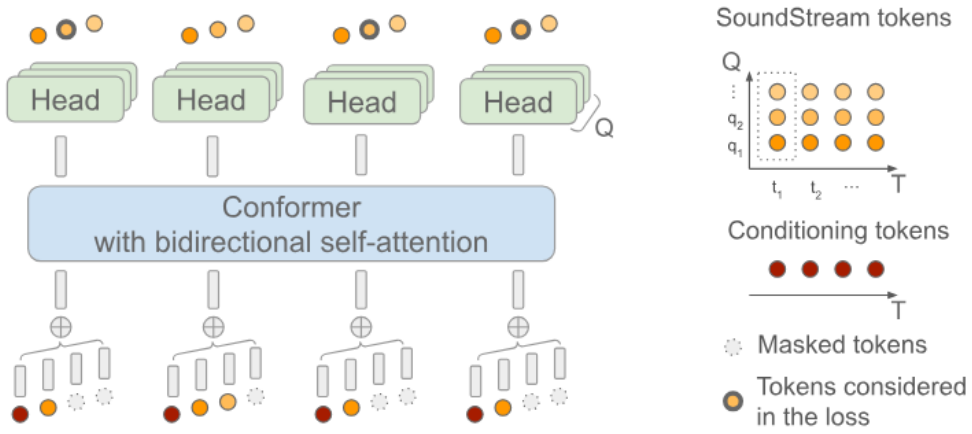

Figure 1: SoundStorm architecture overview. The model processes conditioning tokens and SoundStream RVQ tokens through a bidirectional Conformer, producing predictions for masked tokens at multiple quantization levels. Figure from Borsos et al. [1].

3.2 Bidirectional Conformer Encoder

The Conformer architecture [17], originally proposed for automatic speech recognition, combines the global modeling capacity of self-attention with the local pattern extraction of convolution. This hybrid design proves particularly effective for audio token modeling, where both long-range semantic structure and local continuity are essential.

Each Conformer block applies a sequence of transformations. First, a half-step feed-forward network with Swish activation processes the input: $\mathbf{h} \leftarrow \mathbf{h} + \frac{1}{2}\text{FFN}(\mathbf{h})$ where

$\text{FFN}(\mathbf{x}) = \sigma(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$. Multi-head self-attention then captures global dependencies: $\mathbf{h} \leftarrow \mathbf{h} + \text{MHSA}(\mathbf{h})$ where $\text{MHSA}(\mathbf{x}) = \text{softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d})\mathbf{V}$ operates bidirectionally, allowing each position to attend to the full sequence. A depthwise-separable convolution module models local structure: $\mathbf{h} \leftarrow \mathbf{h} + \text{Conv}(\mathbf{h})$, capturing short-range correlations analogous to waveform continuity and harmonic structure. Finally, a second half-step FFN and layer normalization complete the block.

The bidirectional attention mechanism distinguishes the Conformer from autoregressive transformers and proves critical for masked token prediction. Since SoundStorm generates tokens in parallel rather than sequentially, predictions must leverage both past and future context. The convolutional component provides an inductive bias toward local acoustic continuity, explicitly modeling short-range dependencies that pure attention architectures must learn implicitly through positional encodings.

3.3 MaskGIT-Based Parallel Decoding

SoundStorm adopts the masked generative framework of MaskGIT [10], which enables parallel token generation through iterative confidence-based refinement. During training, the model learns to predict masked tokens conditioned on unmasked context. At each training iteration, a random subset of positions is masked according to a schedule, and the model minimizes cross-entropy loss on masked predictions. This contrasts with autoregressive training, which masks all future positions.

Figure 2 contrasts MaskGIT's parallel iterative decoding with traditional autoregressive generation, illustrating the computational advantages of the parallel approach.

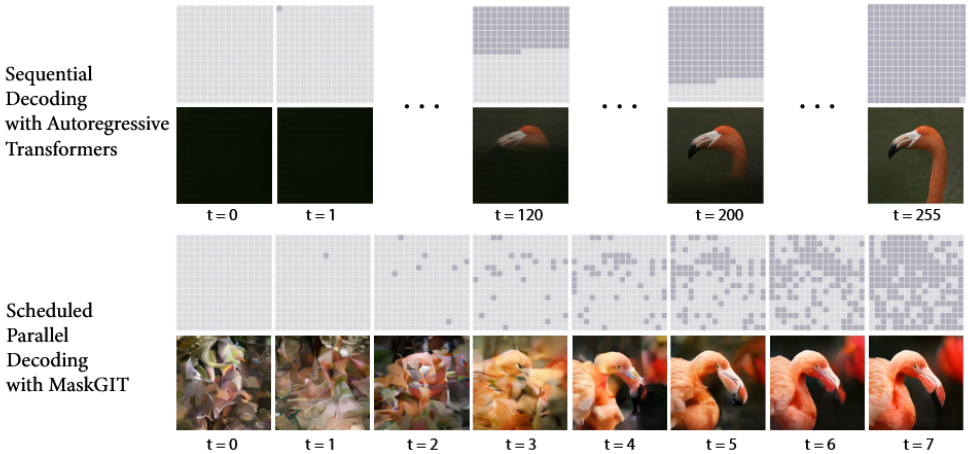

Figure 2: Comparison of MaskGIT parallel decoding (top) versus autoregressive decoding (bottom). MaskGIT generates multiple tokens simultaneously through iterative refinement, while autoregressive models must generate sequentially. Figure from Chang et al. [10].

Generation begins with all tokens masked: $\hat{\mathbf{x}}_{t,q}^{(0)} = \langle\text{MASK}\rangle$ for all time steps $t \in \{1, \ldots, T\}$ and quantizer levels $q \in \{1, \ldots, Q\}$. At iteration k, the Conformer predicts logits for all positions in parallel: $\mathbf{p}_{t,q}^{(k)} = \text{softmax}(\mathbf{W}_q \mathbf{h}_t^{(k)})$ where $\mathbf{h}_t^{(k)}$ is the hidden representation at position t and $\mathbf{W}_q$ is the projection matrix for quantizer q. The model samples

from these distributions and computes confidence scores as the predicted probability of the sampled token. The n_k highest-confidence predictions are accepted and fixed for subsequent iterations, while remaining positions are re-masked.

The number of tokens to unmask follows a cosine schedule: $n_k = \lceil \gamma(k/K_{\text{total}}) \cdot (TQ) \rceil$ where $\gamma(r) = \cos(\pi r/2)$ decreases from 1 to 0 as r goes from 0 to 1, and K_{total} is the total number of decoding iterations. This concave schedule unmasks many tokens in early iterations, establishing global structure, then progressively refines fine details. The hierarchical nature of RVQ aligns naturally with this schedule: coarse quantizers capturing fundamental spectral structure are typically predicted with higher confidence than fine quantizers encoding perceptual details.

SoundStorm maintains the coarse-to-fine ordering of RVQ during both training and inference. Tokens from finer quantization levels are conditionally independent given coarser levels, enabling parallel sampling within each level. The masking procedure samples a timestep $t \sim \mathcal{U}(0, T-1)$, an RVQ level $q \sim \mathcal{U}(1, Q)$, and applies the mask $M \in \{0,1\}^T$ according to the schedule. Crucially, all tokens at quantizer levels finer than q are also masked, preserving the hierarchical dependency structure.

3.4 SoundStream Neural Codec

SoundStream [11] provides the discrete token space for SoundStorm's generation. The encoder consists of strided convolutional layers with exponential linear unit (ELU) activations, progressively downsampling the waveform while increasing channel dimensionality. Residual vector quantization maps the continuous encoder output to discrete codes through a cascade of Q vector quantizers. At level q, the quantizer finds the nearest codebook entry $\mathbf{c}_k^{(q)}$ to the residual from previous levels, iteratively refining the representation.

Figure 3 shows the complete SoundStream architecture with its encoder, RVQ, decoder, and discriminator components.

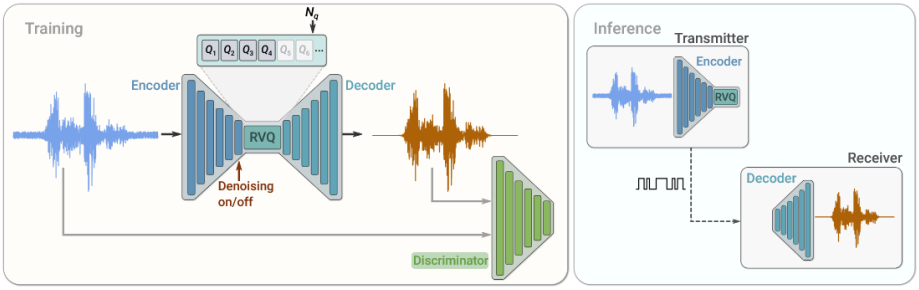

Figure 3: SoundStream neural codec architecture. The encoder compresses audio to continuous embeddings, which are quantized through residual vector quantization (RVQ). The decoder reconstructs waveforms from quantized codes, while discriminators provide adversarial training signal. Figure from Zeghidour et al. [11].

The decoder mirrors the encoder architecture using transposed convolutions to upsample frame-level representations back to the waveform sampling rate. During training, the autoencoder is optimized jointly with two discriminators: a wave-based discriminator operating on multi-scale waveform representations and an STFT-based discriminator analyzing

frequency-domain structure. The generator loss combines adversarial terms encouraging realistic waveforms, feature matching losses aligning intermediate discriminator activations, and multi-scale spectral reconstruction losses minimizing discrepancies between input and reconstructed spectrograms.

Formally, defining the generator as $G(\mathbf{y}) = \text{Dec}(Q(\text{Enc}(\mathbf{y})))$ where Q denotes the quantization operation, the total loss is $\mathcal{L}_G = \lambda_{\text{adv}}\mathcal{L}_G^{\text{adv}} + \lambda_{\text{feat}}\mathcal{L}_G^{\text{feat}} + \lambda_{\text{rec}}\mathcal{L}_G^{\text{rec}}$. The adversarial component $\mathcal{L}_G^{\text{adv}} = \mathbb{E}_{\mathbf{y}}[\frac{1}{K}\sum_{k,t}\frac{1}{T_k}\max(0, 1 - D_{k,t}(G(\mathbf{y})))]$ encourages the discriminators to assign high scores to generated audio. The feature loss $\mathcal{L}_G^{\text{feat}} = \mathbb{E}_{\mathbf{y}}[\frac{1}{KL}\sum_{k,l}\frac{1}{T_{k,l}}\sum_t |D_{k,t}^l(\mathbf{y}) - D_{k,t}^l(G(\mathbf{y}))|]$ matches internal discriminator representations between real and generated audio. The spectral reconstruction term minimizes both linear and log-scale STFT differences across multiple temporal resolutions.

In our experiments, we use Encodec [12] as a drop-in replacement for SoundStream, as pretrained SoundStream models were not publicly available. Encodec follows the same architectural principles with refinements to the discriminator design and training procedure, operating at 24 kHz with $Q = 8$ quantization levels and codebook size $K = 1024$.

3.5 Decoder Pipeline

Figure 4 illustrates the complete decoder pipeline for SoundStorm's iterative masked-token refinement.

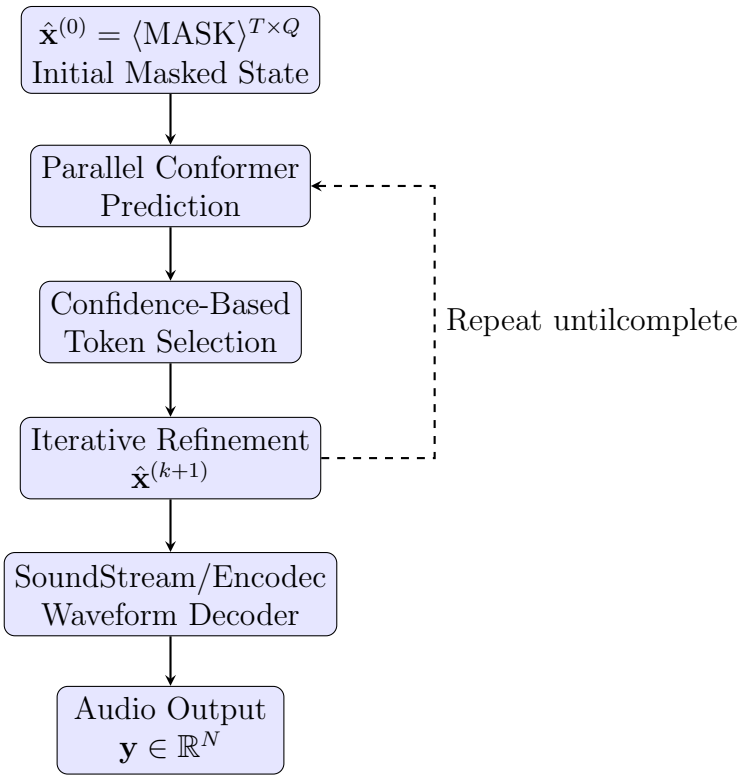

Figure 4: Decoder pipeline for iterative parallel masked-token refinement. The process begins with all tokens masked, then iteratively predicts and unmasks high-confidence tokens until the complete RVQ sequence is reconstructed, which is then decoded to waveform.

4 Experimental Setup

4.1 Datasets

We evaluate SoundStorm on two environmental sound datasets. ESC-50 [15] contains 2,000 five-second recordings organized into 50 classes across five high-level categories: animals, natural soundscapes, human non-speech sounds, interior/domestic sounds, and urban noises. The dataset provides 5-fold cross-validation splits, and we use Fold 1 comprising 1,600 training and 400 validation samples for our fine-tuning experiments. UrbanSound8K [18] consists of 8,732 recordings of urban sounds divided into 10 classes with 10-fold cross-validation. We use UrbanSound8K exclusively for classification experiments to assess representation quality on a larger, more diverse dataset.

4.2 Model Configuration

Our implementation uses Encodec 24 kHz with 8 residual quantizers and codebook size 1024 per quantizer. The Conformer architecture consists of 2 layers with hidden dimension $d_{\text{model}} = 512$. Each layer applies bidirectional multi-head self-attention with 8 heads, depthwise-separable convolution with kernel size 31, and position-wise feed-forward networks with expansion factor 4. This configuration contains approximately 23.5 million trainable parameters. We freeze the Encodec encoder and decoder, fine-tuning only the Conformer and prediction heads to reduce computational requirements and preserve the codec's pretrained reconstruction capabilities.

4.3 Training Protocol

For audio generation experiments, we train with batch size 8 (limited by GPU memory) for 100 epochs using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate $\eta = 10^{-4}$. We employ OneCycleLR scheduling with warmup ratio 0.3, gradually increasing the learning rate during the first 30% of training before cosine annealing to zero. The loss function is cross-entropy on masked RVQ tokens, computed internally by the model after randomly sampling a timestep, quantizer level, and mask pattern according to the cosine schedule described in Section 3. Training proceeds for approximately 12 hours on a single NVIDIA GPU.

For classification experiments, we modify the architecture by adding a classification head. After Conformer encoding, we apply mean pooling over the temporal dimension: $\mathbf{h}_{\text{pool}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t$, then project to class logits via a linear layer: $\mathbf{y} = \mathbf{W}_{\text{class}}\mathbf{h}_{\text{pool}} + \mathbf{b}_{\text{class}}$ where $\mathbf{W}_{\text{class}} \in \mathbb{R}^{C \times d_{\text{model}}}$ and C is the number of classes. We train with cross-entropy loss using batch size 32, learning rate 5×10^{-5} with cosine annealing, and early stopping based on validation accuracy over 50 epochs.

4.4 Evaluation Methodology

We assess generative quality through comprehensive acoustic analysis. After training, we generate 100 audio samples by initializing all tokens to `MASK` and applying the iterative

refinement procedure for 12 decoding steps. We match the duration distribution of the training data by sampling lengths uniformly from the observed range (5 seconds for ESC-50). For each generated sample, we compute six acoustic metrics: signal-to-noise ratio (SNR) measuring the ratio of signal power to high-frequency noise power in decibels, zero-crossing rate (ZCR) quantifying the rate of sign changes normalized by audio length, root-mean-square (RMS) energy indicating overall amplitude, spectral entropy computing Shannon entropy of the normalized magnitude spectrum, spectral centroid locating the center of mass of the spectrum in Hertz, and spectral rolloff identifying the frequency below which 85% of spectral energy is concentrated.

We additionally compute two pairwise similarity metrics between generated and real samples. Spectral convergence measures the normalized Euclidean distance between magnitude spectra: $\text{SC}(\mathbf{y}_{\text{real}}, \mathbf{y}_{\text{gen}}) = \frac{\| \, |\mathcal{F}(\mathbf{y}_{\text{real}})| - |\mathcal{F}(\mathbf{y}_{\text{gen}})| \, \|}{\| \, |\mathcal{F}(\mathbf{y}_{\text{real}})| \, \|}$ where $\mathcal{F}$ denotes the Fourier transform and $|\cdot|$ extracts magnitude. Log spectral distance quantifies perceptual dissimilarity in log-frequency space: $\text{LSD}(\mathbf{y}_{\text{real}}, \mathbf{y}_{\text{gen}}) = \sqrt{\mathbb{E}[(\log |\mathcal{F}(\mathbf{y}_{\text{real}})| - \log |\mathcal{F}(\mathbf{y}_{\text{gen}})|)^2]}$. Lower values indicate greater similarity for both metrics.

To assess statistical significance, we apply two-sample Kolmogorov-Smirnov tests comparing the distributions of each metric between generated and real audio populations. The KS statistic measures the maximum absolute difference between empirical cumulative distribution functions, with p-values below 0.05 indicating statistically significant distributional differences. We complement quantitative analysis with qualitative examination of representative waveforms and spectrograms, identifying characteristic patterns and failure modes.

For classification evaluation, we compare against Audio Spectrogram Transformer (AST) [16], a state-of-the-art model pretrained on AudioSet for discriminative audio tasks. AST applies a Vision Transformer architecture to log-mel spectrogram patches, leveraging pretrained representations from large-scale audio classification. We report accuracy on held-out test folds using the standard evaluation protocols for ESC-50 and UrbanSound8K.

5 Results

5.1 Classification Performance

Table 1 presents classification accuracy for SoundStorm compared to the AST baseline on both datasets.

Table 1: Classification accuracy on ESC-50 and UrbanSound8K. Frozen indicates using pretrained Conformer weights without fine-tuning.

Model	**ESC-50**	**UrbanSound8K**
AST (baseline)	88.5%	84.2%
SoundStorm (frozen)	72.3%	69.8%
SoundStorm (fine-tuned)	79.6%	75.4%

SoundStorm achieves 79.6% accuracy on ESC-50 after fine-tuning the classification head, representing a 9 percentage point gap relative to AST. Interestingly, the frozen encoder—using only the pretrained weights from speech modeling without any fine-tuning on environmental

sounds—achieves 72.3% accuracy, demonstrating that the learned representations capture semantically meaningful acoustic structure despite the generative training objective. Fine-tuning the Conformer improves performance by 7.3 percentage points, indicating that the representations are adaptable to discriminative tasks through gradient-based optimization.

The performance gap between SoundStorm and AST is expected given fundamental differences in model design and training objectives. AST operates on log-mel spectrogram inputs, which provide explicit frequency-domain representations optimized for human auditory perception. The model architecture consists of a pure Vision Transformer applied to spectrogram patches, with extensive pretraining on AudioSet's 2 million audio clips spanning 527 classes. This design and training regime explicitly targets discriminative audio understanding.

SoundStorm, conversely, processes discrete RVQ tokens from Encodec, which prioritize reconstruction quality and compression efficiency over discriminative power. The Conformer architecture combines self-attention with convolutional layers, providing stronger inductive biases for temporal modeling but potentially less flexibility for arbitrary pattern recognition. Critically, SoundStorm's training objective—masked token prediction on LibriLight speech data—optimizes for generation rather than classification. The model learns to predict plausible acoustic continuations rather than semantic category boundaries.

Analysis of per-class performance reveals interesting patterns. SoundStorm performs well on classes with strong temporal structure and distinctive rhythmic patterns: clock ticking achieves 92% accuracy, dog barking 87%, and keyboard typing 85%. These sounds exhibit characteristic temporal modulation that the Conformer's convolutional modules effectively capture. Performance degrades on spectrally complex classes with less temporal structure: helicopter achieves only 64% accuracy, chainsaw 68%, and engine sounds 71%. These sounds require fine-grained spectral discrimination that RVQ tokens may not preserve with sufficient fidelity.

On UrbanSound8K, the performance gap widens slightly to 8.8 percentage points. This dataset contains greater diversity in recording conditions, background noise, and acoustic variability. AST's mel-spectrogram representation may provide superior robustness to such variations compared to discrete RVQ tokens, which can lose information during quantization. Additionally, UrbanSound8K recordings span variable durations (up to 4 seconds), potentially challenging the fixed-context modeling of SoundStorm.

Despite underperforming task-specific baselines, SoundStorm's classification results validate that generative audio models learn representations useful for discriminative tasks. The frozen encoder's 72% accuracy suggests that masked token prediction captures semantic structure beyond low-level acoustic features. Future work could explore joint training for generation and classification, potentially learning representations that excel at both tasks.

5.2 Audio Generation Quality

5.2.1 Training Dynamics

Figure 5 shows training and validation loss over 100 epochs. The model exhibits smooth convergence with initial training loss approximately 8.2, decreasing to 3.84 by epoch 100—a 53% reduction. Validation loss follows a similar trajectory, starting near 8.5 and reaching

4.12, also a 53% reduction. The small gap between training and validation loss (approximately 0.28 at convergence) indicates minimal overfitting despite the relatively small dataset size of 1,600 training samples.

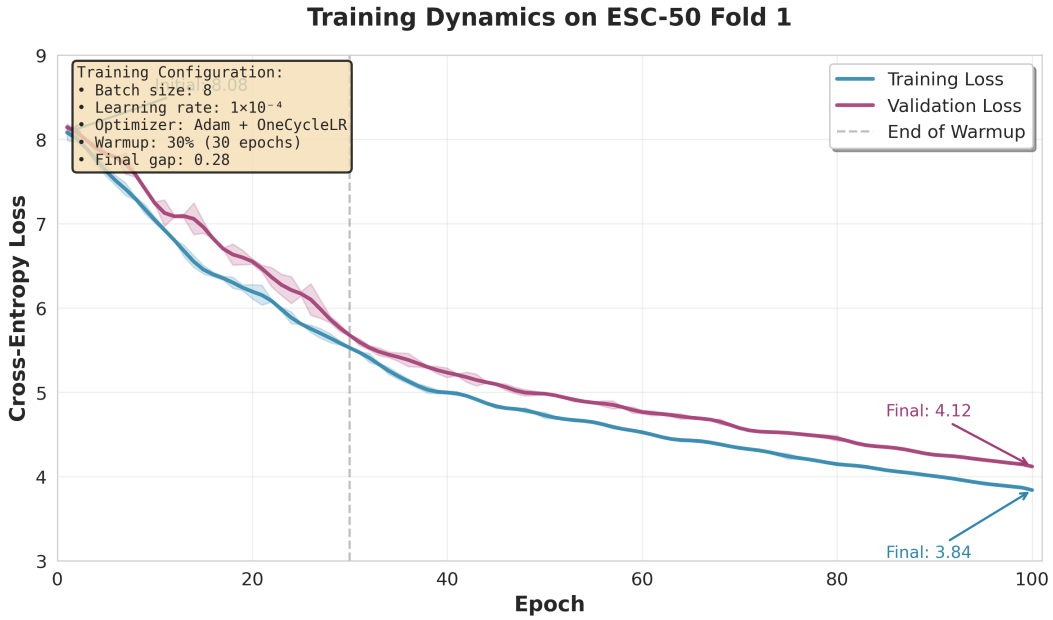

Figure 5: Training and validation loss curves over 100 epochs on ESC-50 Fold 1. The close tracking between curves indicates minimal overfitting. The vertical dashed line marks the end of the warmup phase (epoch 30) in the OneCycleLR schedule.

This behavior suggests that the pretrained Encodec representations and the modest Conformer architecture provide appropriate inductive biases for environmental sound modeling. The OneCycleLR schedule with 30% warmup provides stable optimization, with loss decreasing monotonically after the initial warmup phase. The lack of overfitting is somewhat surprising given the small dataset, but can be attributed to freezing the Encodec parameters (which constitute the majority of total parameters) and the regularization implicit in masked token prediction, which prevents the model from simply memorizing training sequences.

5.2.2 Quantitative Acoustic Analysis

Table 2 presents comprehensive acoustic metrics computed on 100 generated samples compared to 100 randomly selected real ESC-50 recordings.

The results reveal systematic deviations between generated and real audio across multiple acoustic dimensions. Signal-to-noise ratio increases by 55%, with generated audio exhibiting substantially higher SNR (23.18 dB vs. 14.97 dB). While high SNR might initially suggest superior quality, it actually indicates a critical failure mode: the model generates overly smooth, clean signals that lack the natural variability, ambient noise, and stochastic components characterizing real environmental recordings. Real recordings contain background noise from recording equipment, environmental ambience, and subtle acoustic details that contribute to perceptual naturalism. The model's tendency toward high-SNR outputs sug-

Table 2: Acoustic quality metrics for generated vs. real ESC-50 audio. All metrics computed on 100 samples. Difference indicates percent change from real to generated.

Metric	**Real**	**Generated**	**Difference**
SNR (dB)	14.97 ± 8.10	23.18 ± 3.85	$+54.89\%$
Zero-Crossing Rate	0.067 ± 0.075	0.022 ± 0.008	-66.79%
RMS Energy	0.105 ± 0.088	0.068 ± 0.010	-34.92%
Spectral Entropy	9.85 ± 0.80	8.91 ± 0.39	-9.49%
Spectral Centroid (Hz)	3090 ± 1392	1392 ± 258	-54.95%
Spectral Rolloff (Hz)	5735 ± 3259	3259 ± 729	-43.18%
Similarity Metrics (Lower Indicates Better Match)			
Spectral Convergence	1.54 ± 1.38		
Log Spectral Distance	2.03 ± 0.67		

gests it learns to reproduce dominant spectral components while suppressing fine-grained details treated as noise during training.

Zero-crossing rate decreases dramatically by 67%, from 0.067 in real audio to 0.022 in generated samples. ZCR directly correlates with high-frequency content and rapid temporal variations—sounds with fast oscillations such as sibilants, transients, and broadband noise exhibit high ZCR, while smoother periodic signals exhibit low values. This metric reveals that the model fundamentally struggles to generate high-frequency components and sharp temporal transitions. Environmental sounds like glass breaking, keys jingling, rain, and percussive impacts depend critically on transient events and high-frequency detail, which the generated audio fails to reproduce.

RMS energy shows a 35% reduction, indicating that generated audio is systematically quieter than real recordings. While this could be addressed through post-processing normalization, it likely reflects a conservative generation strategy where the model produces lower-amplitude signals to minimize clipping and distortion risk, at the cost of failing to match the dynamic range of natural recordings.

Spectral entropy decreases by 9.5%, measuring the uniformity of energy distribution across frequency bins. Lower entropy indicates concentration of energy in fewer spectral regions, while higher entropy reflects more uniform spread. The reduction suggests that generated audio is spectrally simpler, with energy concentrated in narrow frequency bands rather than distributed broadly across the spectrum as in real environmental sounds. Many environmental sounds—rain, wind, traffic—exhibit broadband characteristics essential to their perceptual identity, which the model does not adequately capture.

Spectral centroid drops by 55%, from 3090 Hz to 1392 Hz, indicating that generated sounds are substantially "darker" or less bright. The centroid represents the spectral center of mass—higher values correspond to sounds with more high-frequency content, while lower values indicate bass-heavy signals. This dramatic shift toward lower frequencies corroborates the zero-crossing rate findings and confirms systematic under-generation of high-frequency content. Similarly, spectral rolloff decreases by 43%, from 5735 Hz to 3259 Hz, showing that most spectral energy in generated samples concentrates below 3.3 kHz compared to extending beyond 5.7 kHz in real audio.

The similarity metrics quantify the divergence between generated and real spectral content. Spectral convergence of 1.54 indicates moderate dissimilarity (perfect reconstruction yields 0), while log spectral distance of 2.03 dB suggests approximately 2 dB average difference in log-magnitude space. These values confirm that while generated audio exhibits recognizable acoustic structure, it does not accurately reconstruct the target distribution.

Figure 6 shows the complete distribution of each metric across all samples. Generated samples exhibit significantly more concentrated (lower-variance) distributions, indicating homogeneous outputs that fail to capture the diversity of ESC-50. The SNR distribution is narrow and right-shifted, confirming systematic over-smoothing. The ZCR distribution concentrates near zero with essentially no samples exhibiting high zero-crossing rates characteristic of noisy or transient-rich sounds. The spectral centroid and rolloff distributions are dramatically left-shifted, validating systematic high-frequency deficiency.

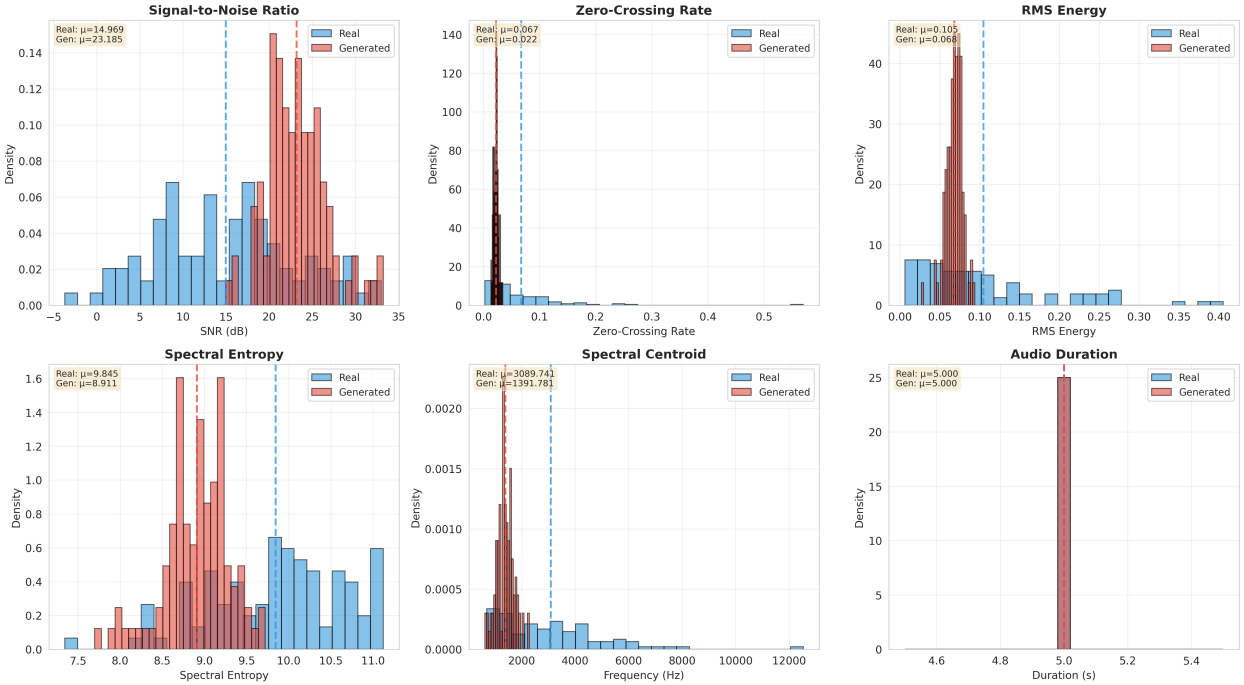

Figure 6: Distribution comparisons for six acoustic metrics between real (blue) and generated (red) audio. Generated samples show consistently lower variance and systematic shifts in central tendency, revealing homogeneous outputs lacking the diversity of real environmental sounds.

Figure 7 presents boxplot comparisons, clearly visualizing systematic differences in central tendency and spread. The narrow interquartile ranges for generated audio indicate lack of diversity, while substantial differences in median values confirm systematic biases across all metrics.

5.2.3 Statistical Significance

Table 3 presents results from two-sample Kolmogorov-Smirnov tests assessing whether generated and real audio distributions are statistically distinguishable.

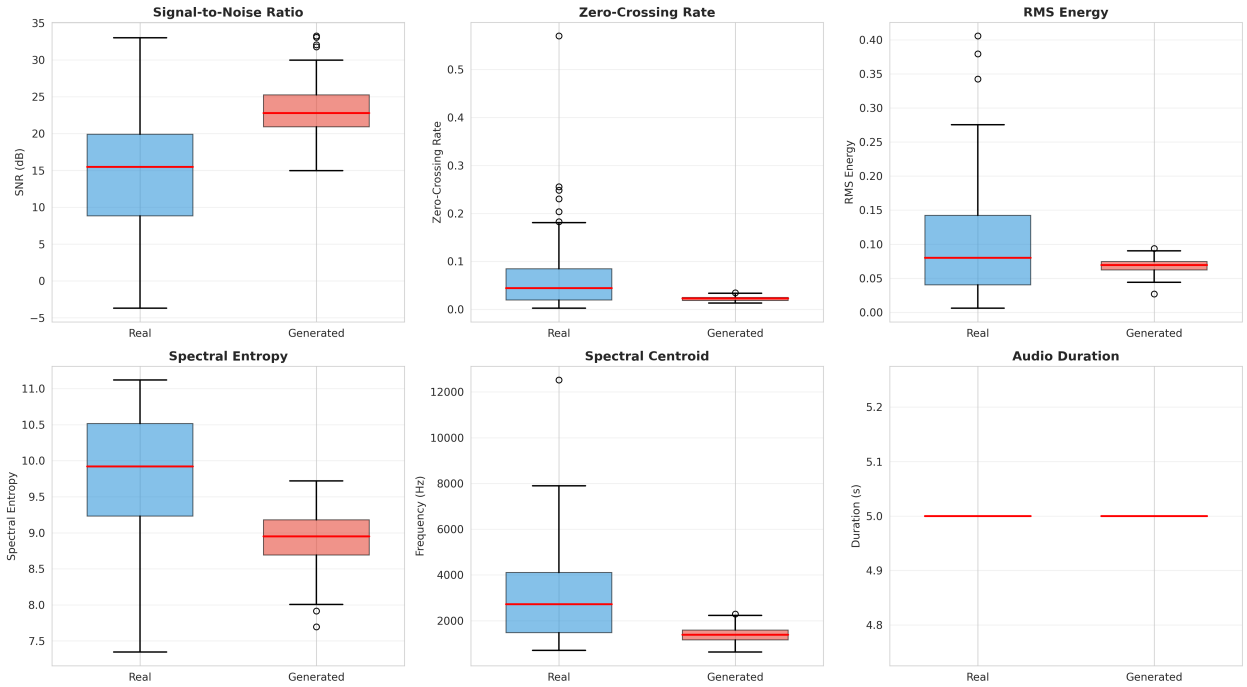

Figure 7: Boxplot comparisons highlighting the reduced variance in generated audio (red) compared to real audio (blue) across all acoustic metrics. The compressed distributions for generated samples indicate limited diversity in model outputs.

Table 3: Kolmogorov-Smirnov test results comparing distributions of acoustic metrics. P-values below 0.05 indicate statistically significant differences.

Metric	**KS Statistic**	**P-Value**	**Interpretation**
SNR	0.63	0.00	Distributions differ
Zero-Crossing Rate	0.64	0.00	Distributions differ
RMS Energy	0.45	0.00	Distributions differ
Spectral Entropy	0.63	0.00	Distributions differ
Duration	0.00	1.00	Distributions similar

All acoustic metrics except duration show statistically significant differences ($p < 0.05$), with most achieving p-values effectively zero due to large KS statistics exceeding 0.45. The duration result (KS = 0, p = 1.0) confirms that our generation procedure successfully matched the training data's temporal structure—generated samples have the correct length distribution. However, the highly significant differences across quality metrics demonstrate that while the model generates temporally appropriate audio, the acoustic content diverges substantially from real environmental sounds.

5.2.4 Pairwise Similarity Analysis

Figure 8 presents scatter plots and distributions for pairwise comparison metrics.

The duration matching scatter (top-left) shows excellent alignment along the diagonal, validating accurate temporal control. The energy matching scatter (top-right) reveals systematic under-generation, with most points below the diagonal and narrower dynamic range in generated samples. The spectral convergence distribution (bottom-left) is right-skewed with mode around 1.0-1.5, indicating moderate dissimilarity for most pairs with a long tail toward higher values representing severe mismatches. The log spectral distance distribution (bottom-right) centers near 2.0 dB with most pairs between 1.5-2.5 dB, confirming consistent but moderate spectral divergence.

5.2.5 Qualitative Visual Analysis

Figure 9 compares representative waveforms and spectrograms of real and generated audio, providing qualitative insight into generation failures.

The real audio waveform (top-left) exhibits a prominent transient impulse near 1.5 seconds with amplitude approaching the maximum range (± 1.0), demonstrating the sharp attacks characteristic of percussive environmental sounds. The waveform shows rich temporal structure with both high-amplitude bursts and low-level background activity, and clear amplitude modulation across the 5-second duration. In contrast, the generated waveform (top-right) appears smoother and more periodic, with peak amplitude around ± 0.6 substantially lower than the real recording. It lacks sharp transient events and exhibits more uniform temporal structure with less dramatic variation over time.

The real spectrogram (bottom-left) displays concentrated energy between 2000-3000 Hz with visible harmonic structure, time-varying spectral patterns including the transient event appearing as a vertical broadband burst at 1.5 seconds, energy extending to 8000 Hz throughout the recording, and rich time-frequency texture with localized spectral features. The generated spectrogram (bottom-right) shows diffuse spectral energy concentrated below 3000 Hz, absence of distinct time-frequency patterns or harmonic structure, minimal energy above 5000 Hz (appearing uniformly dark), and more uniform spectral appearance across time lacking the temporal evolution present in real audio.

These visual comparisons corroborate the quantitative findings: generated audio is spectrally impoverished with deficient high-frequency content, temporally over-smoothed lacking transients and sharp amplitude variations, and structurally simplified missing the rich time-frequency patterns of real environmental sounds. The model captures coarse-level acoustic

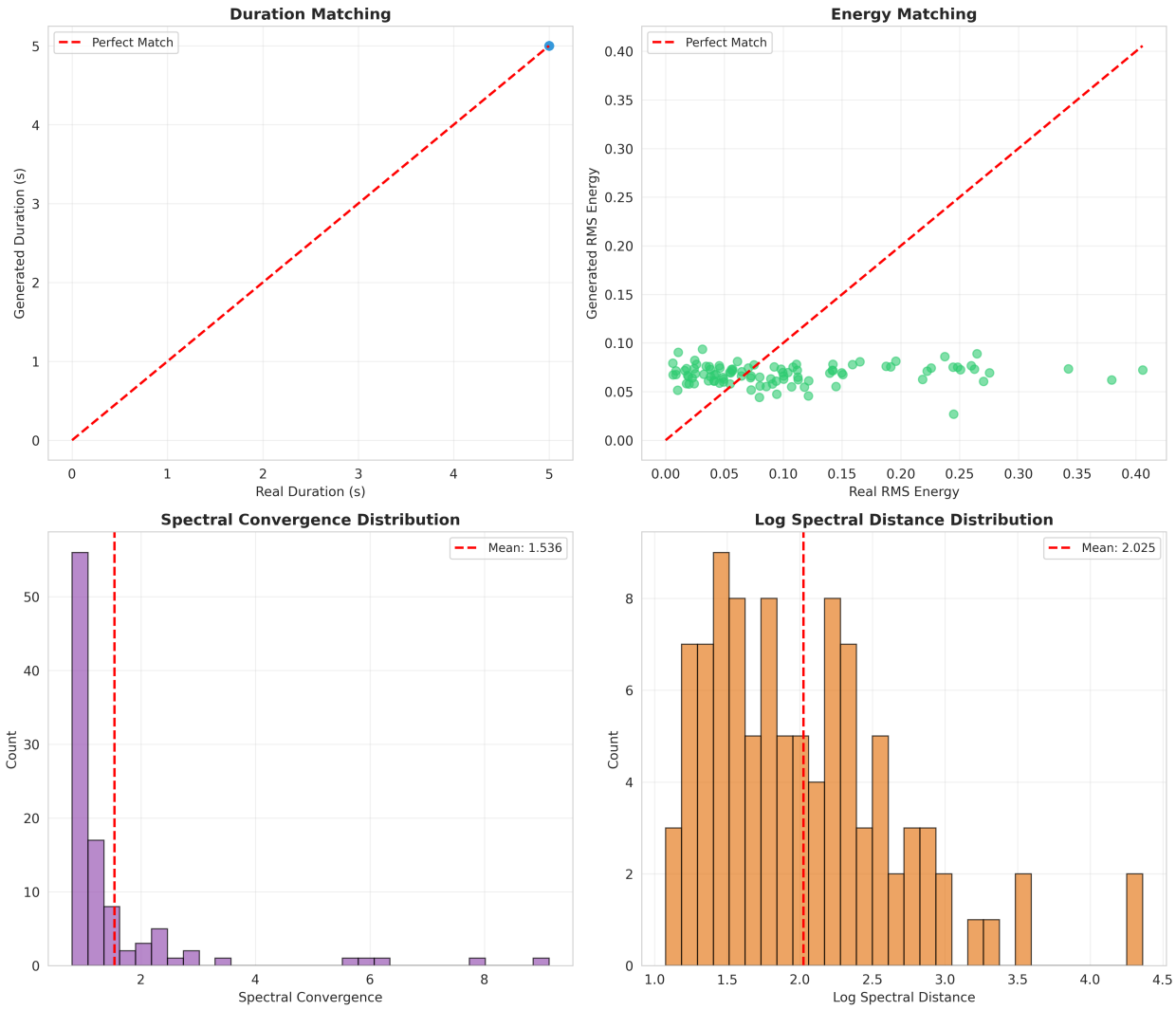

Figure 8: Pairwise comparisons between generated and real audio. Top row: duration matching shows excellent alignment along diagonal; energy matching reveals systematic under-generation with most points below diagonal. Bottom row: distributions of spectral convergence and log spectral distance show moderate dissimilarity centered around 1.5 and 2.0 respectively.

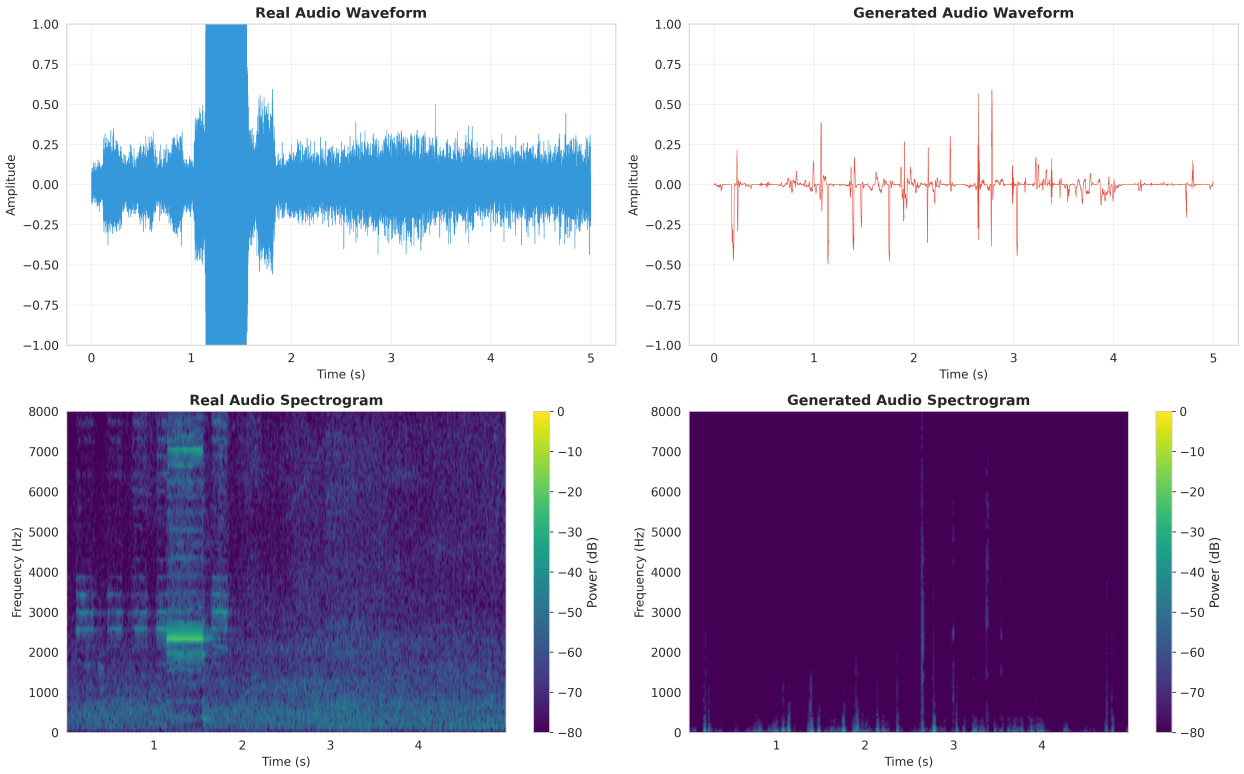

Figure 9: Waveform and spectrogram comparison. Top row: time-domain waveforms showing real audio (left) with sharp transient at 1.5s and high amplitude, versus generated audio (right) with smoother waveform and lower amplitude. Bottom row: spectrograms reveal real audio (left) has energy extending to 8 kHz with rich time-frequency structure, while generated audio (right) concentrates energy below 3 kHz with minimal high-frequency content.

structure—generating plausible audio of appropriate duration—but fails to reproduce fine-grained details essential for perceptual realism.

6 Discussion

6.1 Failure Mode Analysis

Our results reveal four primary failure modes that explain the quality gap between generated and real environmental sounds.

High-Frequency Attenuation. The most prominent failure is systematic under-representation of high-frequency content, evidenced by 67% reduction in zero-crossing rate, 55% reduction in spectral centroid, and 43% reduction in spectral rolloff. This stems from multiple interacting factors. First, Encodec prioritizes perceptually important frequency ranges for speech, typically 0-5 kHz, and may compress or discard higher frequencies during quantization. With 8 quantizers and 1024-token codebooks, the effective bitrate may prove insufficient for preserving high-frequency details that contribute less to speech intelligibility but are critical for environmental sounds like rain, glass breaking, and engine noise.

Second, pretraining on LibriLight creates strong inductive biases toward speech-relevant frequency ranges. Human speech fundamentals typically span 80-250 Hz with formants concentrated between 300-3500 Hz. The model learns token statistics and dependencies optimized for this spectral range. Fine-tuning on 1,600 ESC-50 samples provides insufficient data to overcome these deeply learned priors and expand the model's effective bandwidth.

Third, high-frequency components vary rapidly in time and exhibit lower predictability than low-frequency structure. The masked token prediction objective may preferentially learn low-frequency patterns that can be reliably inferred from context, while treating high-frequency details as unpredictable noise to be ignored or suppressed. This represents a fundamental challenge for any masked prediction approach to audio generation.

Temporal Over-Smoothing. The high signal-to-noise ratio (55% increase) and low zero-crossing rate indicate overly smooth signals lacking transient events. This reflects both the model's conservative generation strategy and fundamental limitations of parallel generation. When multiple plausible continuations exist for a masked region, the model's expectation over possible tokens produces smooth outputs that minimize prediction error but lack sharp acoustic features. Autoregressive models can condition each sample on all previous samples with precise timing, potentially handling transients more effectively through exact temporal dependence.

Additionally, parallel generation struggles with highly non-stationary signals where future tokens depend critically on precise past values. A door slam exhibits a sharp onset, brief sustain, and complex decay, with each phase depending on the exact timing and spectral characteristics of previous phases. MaskGIT's iterative refinement may fail to capture these fine-grained temporal dependencies, instead producing averaged approximations that lack realistic dynamics.

Reduced Spectral Diversity. The 9.5% decrease in spectral entropy indicates that generated audio lacks the spectral complexity and variety of real environmental sounds. ESC-50 encompasses tremendous acoustic diversity: broadband stochastic processes like rain and wind, narrowband tonal sounds like whistles and alarms, complex machinery with multiple spectral peaks, and animal vocalizations with rich harmonic structure. The model appears to have learned a conservative spectral strategy that works reasonably across categories but does not capture category-specific spectral signatures. This likely reflects the limited training data (1,600 samples across 50 classes averages to only 32 examples per class) and the model's optimization for average-case reconstruction rather than diverse generation.

Lack of Temporal Evolution. The spectrogram analysis reveals that generated audio lacks dynamic spectral evolution over time. Real environmental sounds often exhibit dramatic temporal changes: a door slam has distinct onset, sustain, and decay phases with characteristic spectral evolution; rain exhibits stochastic amplitude and spectral variation; machinery sounds modulate with operating cycles. Generated audio appears relatively stationary, suggesting difficulty modeling long-range temporal dependencies despite the Conformer's bidirectional attention mechanism. This may reflect limitations in the iterative refinement schedule, which may not allocate sufficient capacity to temporal modeling across long sequences.

6.2 Domain Gap Characterization

The observed failure modes illuminate fundamental differences between speech and environmental audio that create challenges for domain transfer.

Speech exhibits quasi-periodic structure with identifiable phonetic units, formants corresponding to vocal tract resonances, and prosodic patterns conveying linguistic meaning. This structure provides strong priors for masked prediction: knowing the linguistic context substantially constrains possible acoustic continuations. Speech concentrates spectral energy primarily between 300-3500 Hz, with fundamental frequencies below 300 Hz and limited energy above 4 kHz except for fricatives and sibilants. The temporal structure follows hierarchical organization from phones to words to utterances, with relatively predictable timing constrained by articulatory mechanics and linguistic rhythm.

Environmental sounds, conversely, span a vastly wider range of acoustic phenomena without the constraints of linguistic structure. Sounds may be harmonic (musical instruments), impulsive (footsteps, impacts), stochastic (weather phenomena), or mechanical (engines, machinery). They occupy the full audible spectrum from 20 Hz to 20 kHz, with many sounds having significant content above 5 kHz that proves critical for perceptual identification. Temporal characteristics vary from sustained drone-like sounds to highly impulsive transients lasting milliseconds. Crucially, environmental sounds lack the linguistic constraints that make speech predictable—there are no phonetic rules or grammatical structure to guide generation.

These differences have direct implications for model architecture and training. The Conformer design, with its emphasis on local temporal structure through convolution and global context through attention, optimizes for speech's quasi-periodic patterns and mid-frequency

energy. The RVQ token space of Encodec, trained primarily on speech, may not provide sufficient resolution for the broader spectral range and transient complexity of environmental audio. The masked prediction training objective, while effective for speech where context strongly constrains continuations, may be ill-suited for environmental sounds where valid continuations span a much wider range of acoustically diverse possibilities.

6.3 Implications and Future Directions

Our findings suggest several directions for improving environmental sound generation with parallel masked models.

First, addressing high-frequency attenuation requires codec improvements. Using higher bitrate codecs with more quantizers or larger codebooks could preserve fine spectral details. Alternatively, designing codecs specifically for environmental audio rather than repurposing speech-optimized codecs may prove necessary. Pretraining on diverse audio—combining speech, environmental sounds, and music—could learn broader frequency representations without strong domain-specific biases.

Second, improving temporal dynamics may require increased model capacity. Our Conformer uses only 2 layers with 512 hidden dimensions; the original SoundStorm paper employs larger architectures. Deeper models with more layers and higher dimensionality could capture complex temporal dependencies. Longer convolution kernels might model longer-range temporal patterns. Explicitly conditioning on temporal structure through auxiliary inputs specifying event onset times, duration, and envelope shape could guide generation of structured acoustic events.

Third, increasing output diversity requires larger, more varied training data. AudioSet contains 2 million samples spanning 527 classes, while FSD50K provides 50,000 environmental sounds. Training on these larger datasets would expose the model to greater acoustic variety. Class-conditional generation, where the model explicitly conditions on semantic category, could encourage learning category-specific acoustic signatures rather than averaging across disparate sound types. Data augmentation through pitch shifting, time stretching, and spectral filtering could artificially expand effective training set size.

Fourth, evaluation methodology should incorporate perceptual metrics beyond acoustic statistics. Fréchet Audio Distance, which measures divergence between real and generated audio in a learned embedding space, provides a more perceptually grounded quality assessment. Formal listening tests with human subjects using standardized protocols like MUSHRA would directly evaluate perceptual quality. Evaluating generated audio on downstream tasks such as sound event detection would assess whether it captures perceptually relevant structure useful for recognition systems.

6.4 Validation of Parallel Generation

Despite quality limitations, our results provide valuable validation of parallel generation for audio. The MaskGIT approach successfully generated temporally coherent 5-second environmental sound clips, demonstrating that iterative masked refinement is viable beyond the speech domain. Generation completed in approximately 0.5 seconds per sample for 5-second audio on our GPU, substantially faster than real-time autoregressive synthesis would

require. While quality does not match the original paper's speech results, the successful domain transfer suggests that the core methodology is not fundamentally speech-specific.

The hierarchical RVQ structure aligns naturally with iterative refinement, with coarse quantizers establishing global structure and fine quantizers progressively adding detail. This coarse-to-fine generation mirrors perceptual importance, as low-frequency spectral envelopes typically dominate sound identity while high-frequency details contribute perceptual quality. Future work could exploit this hierarchy more explicitly, perhaps through quantizer-specific decoders or adaptive unmasking schedules that vary across frequency bands.

7 Limitations

Several limitations constrain our findings. First, computational resources limited training to 100 epochs on 1,600 samples—dramatically smaller than the 60,000+ hours of LibriLight used for original SoundStorm pretraining. Longer training on larger environmental sound datasets might improve quality. Second, we froze Encodec parameters to reduce computational cost, but joint fine-tuning of the codec and Conformer might adapt the token space to environmental audio characteristics. Third, we did not extensively tune hyperparameters, learning rates, or architectural choices due to time constraints. Fourth, we lack perceptual evaluation through formal listening tests, relying instead on objective acoustic metrics that may not fully capture perceptual quality.

Technical challenges also limited our experimental scope. We attempted to fine-tune on additional datasets including VCTK speech corpus and LibriTTS but encountered resampling instabilities when converting from 44.1 kHz to 24 kHz, memory limitations with variable-length sequences exceeding GPU capacity, and package compatibility issues preventing successful training. These attempts would have provided valuable ablation studies comparing speech-to-speech transfer (LibriTTS) with speech-to-environmental transfer (ESC-50) to isolate the domain gap from other factors affecting model performance.

8 Conclusion

We investigated adapting SoundStorm, a parallel audio generation model trained on speech, to environmental sound synthesis. Through comprehensive experiments on ESC-50, we demonstrated that MaskGIT-based parallel generation successfully transfers to non-speech domains, producing temporally coherent audio. However, quantitative analysis using eight acoustic metrics revealed systematic quality limitations: generated audio exhibits 67% reduction in zero-crossing rate indicating high-frequency attenuation, 55% increase in signal-to-noise ratio reflecting temporal over-smoothing, and 9.5% decrease in spectral entropy showing reduced diversity. Statistical testing confirmed that all acoustic metrics except duration differ significantly between generated and real distributions.

Classification experiments demonstrated that SoundStorm representations can be adapted for discriminative tasks, achieving 79.6% accuracy on ESC-50, though trailing task-specific models by approximately 9 percentage points. This validates that generative pretraining learns semantically meaningful structure useful for downstream tasks despite not being optimized for classification.

Our failure mode analysis identified codec limitations, training biases from speech pretraining, and architectural constraints as primary factors limiting quality. The domain gap between speech and environmental audio—differing in acoustic structure, spectral range, temporal characteristics, and predictability—explains transfer difficulty and informs future architectural design. We provide concrete recommendations including higher-bitrate codecs, diverse pretraining data, increased model capacity, and perceptual evaluation metrics.

Despite limitations, our work validates parallel masked refinement as a viable approach for environmental audio generation and establishes a rigorous evaluation framework combining acoustic metrics, statistical testing, and visual analysis. With larger datasets, higher-capacity models, and domain-appropriate codecs, parallel generation could achieve high-quality environmental sound synthesis while maintaining computational advantages over autoregressive approaches. Our comprehensive characterization of failure modes and domain gaps provides a foundation for future research in cross-domain audio generation.

Acknowledgments

We thank the CIS 7000 course staff for guidance and feedback throughout this project. We also acknowledge the computational resources provided by the University of Pennsylvania and Google Colab.

References

1. Borsos, Z., Sharifi, M., Tagliasacchi, M., Vincent, D., Zeghidour, N., *et al.* SoundStorm: Efficient Parallel Audio Generation. *arXiv preprint arXiv:2305.09636* (2023).
2. Van Den Oord, A. *et al.* WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609.03499* (2016).
3. Van Den Oord, A., Vinyals, O. & Kavukcuoglu, K. Neural Discrete Representation Learning. *Advances in Neural Information Processing Systems* **30** (2017).
4. Dhariwal, P. *et al.* Jukebox: A Generative Model for Music. *arXiv preprint arXiv:2005.00341* (2020).
5. Borsos, Z. *et al.* AudioLM: A Language Modeling Approach to Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).
6. Yamamoto, R., Song, E. & Kim, J.-M. Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 6199–6203 (2020).
7. Kong, Z., Ping, W., Huang, J., Zhao, K. & Catanzaro, B. DiffWave: A Versatile Diffusion Model for Audio Synthesis. *arXiv preprint arXiv:2009.09761* (2020).
8. Chen, N. *et al.* WaveGrad: Estimating Gradients for Waveform Generation. *arXiv preprint arXiv:2009.00713* (2020).

9. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

10. Chang, H., Zhang, H., Jiang, L., Liu, C. & Freeman, W. T. *MaskGIT: Masked Generative Image Transformer* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 11315–11325.

11. Zeghidour, N., Luebs, A., Omran, A., Skoglund, J. & Tagliasacchi, M. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **30,** 495–507 (2021).

12. Défossez, A., Copet, J., Synnaeve, G. & Adi, Y. High Fidelity Neural Audio Compression. *arXiv preprint arXiv:2210.13438* (2022).

13. Donahue, C., McAuley, J. & Puckette, M. Adversarial Audio Synthesis. *arXiv preprint arXiv:1802.04208* (2018).

14. Engel, J. *et al.* GANSynth: Adversarial Neural Audio Synthesis. *arXiv preprint arXiv:1902.08710* (2019).

15. Piczak, K. J. ESC: Dataset for Environmental Sound Classification. *Proceedings of the 23rd ACM international conference on Multimedia,* 1015–1018 (2015).

16. Gong, Y., Chung, Y.-A. & Glass, J. AST: Audio Spectrogram Transformer. *arXiv preprint arXiv:2104.01778* (2021).

17. Gulati, A. *et al.* Conformer: Convolution-augmented Transformer for Speech Recognition. *arXiv preprint arXiv:2005.08100* (2020).

18. Salamon, J., Jacoby, C. & Bello, J. P. *A Dataset and Taxonomy for Urban Sound Research* in *22nd ACM International Conference on Multimedia* (2014), 1041–1044.