

SCAPES: SEMANTICALLY CONDITIONED AUTOREGRESSIVE PRIOR FOR ENVIRONMENTAL SOUNDS

Esteban Gutiérrez, Lonce Wyse, Frederic Font and Xavier Serra

Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain

<esteban.gutierrezc, lonce.wyse, frederic.font, xavier.serra>@upf.edu

ABSTRACT

As generative audio models grow in complexity, the computational and ecological costs of synthesizing everyday sounds have become increasingly prohibitive, often requiring industrial-scale resources and massive datasets. In this paper, we present SCAPES: a Semantically Conditioned Autoregressive Prior for Environmental Sounds. SCAPES is a lightweight, resource-efficient generative model designed to synthesize high-fidelity environmental textures through high-level semantic control. By operating on the continuous latent manifold of a neural audio codec, our approach bypasses the rigid structural constraints inherent to discrete tokenization. We propose a segmentation strategy that decomposes audio into overlapping segments, enabling a Continuous Normalizing Flow (CNF) to model the evolution of latent trajectories using Flow Matching. Our experiments demonstrate that a 36-million parameter instance of SCAPES can be trained on limited, uncured datasets using a single consumer-grade GPU. Notably, convergence is achieved after training for approximately twice the source audio duration, yielding high-fidelity outputs with robust long-term stability and semantic consistency. Furthermore, we showcase the model’s capacity for smooth semantic interpolation, providing a flexible and accessible tool for open research and creative sound design. Code, pretrained weights, audio examples, and an interactive demo are publicly available on our project page cordutie.github.io/projects/scapes.html.

1. INTRODUCTION

The concept of “everyday listening” emphasizes the human experience of hearing events in the world rather than the abstract properties of the sounds themselves [1]. While traditional “musical listening” focuses on perceptual dimensions like pitch, loudness, and timbre, everyday listening is primarily concerned with the physical sources, materials, and the underlying semantic meaning of the interactions that produce the audio.

When viewed through this ecological lens, the acoustic domain reveals itself as a vastly diverse field of environmental sounds. This broad spectrum encompasses everything from discrete, structured interactions, such as keyboard strokes, applause, or a helicopter rotor, to continuous, stochastic textures, like rushing rivers, crackling bonfires, wind blowing, or heavy rain.

Copyright: © 2026 Esteban Gutiérrez et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, adaptation, and reproduction in any medium, provided the original author and source are credited.

With the rapid advancement of generative audio models in recent years, there has been a growing research focus on *controllability* [2, 3, 4, 5, 6], defined as the ability to intuitively guide the generative process. Nevertheless, designing intuitive control mechanisms for environmental sounds remains a significant challenge, owing to their diversity and lack of inherent structure. In the context of generative audio models, user-controllable features can be broadly categorized as follows:

1. **Signal properties:** Purely signal-based metrics, such as spectral coefficients, zero-crossing rate, and signal energy.
2. **Perceptual features:** Properties derived from psychoacoustic models, including pitch, loudness, and perceptual summary statistics.
3. **Data-driven representations:** Abstract features learned from the data distribution, such as unconditioned latent spaces.
4. **Source properties:** Low-level physical parameters of the sound-generating mechanism, such as string length or material density.
5. **Semantic features:** High-level, human-intuitive descriptors, such as text prompts or sound event labels.

In the realm of open, lightweight, and controllable sound synthesis, various models have successfully leveraged a subset of these modalities. Pioneering physical modeling approaches [7, 8] are fundamentally driven by *source properties*. Texture synthesis frameworks, such as those introduced by McDermott and Simoncelli [9, 10] alongside TexStat [11], generate audio governed by *perceptual* summary statistics. Differentiable digital signal processing (DDSP) techniques [2, 12] typically combine *signal* and *perceptual* features to guide the generation process. Additionally, alternative texture synthesis models like SynTex [13] rely on varying combinations of *source* and *signal* properties. Finally, modern neural audio models based on Variational Autoencoders, including [14] and other widely recognized architectures such as RAVE [15] and AFTER [3], synthesize sound by navigating *data-driven representations* defined inherently by the model architecture.

Since everyday listening is fundamentally anchored in high-level semantics, *semantic features* offer the most intuitive modality for controlling everyday sound generation. However, their inherent coarseness lacks the fine-grained structural information needed to directly synthesize complex waveforms. Mapping such sparse, high-level descriptors to dense audio signals is a highly under-constrained problem; consequently, successful approaches have largely depended on leveraging immense datasets to learn these complex distributions. As a result, semantic control is currently dominated by large-scale text-to-audio frameworks [16, 17, 18, 4], with a growing emphasis on closed-source, proprietary systems

such as Meta’s Audiobox [19], Adobe’s Sketch2Sound [5], and Nvidia’s Fugatto [20]. Ironically, this reliance on scale implies that synthesizing even simple, natural environmental sounds requires deploying highly resource-intensive models, an approach that carries a substantial ecological footprint.

To address this, we present SCAPES: a Semantically Conditioned Autoregressive Prior for Environmental Sounds. SCAPES is a generative model trained without manual annotations, designed for high-fidelity, semantically controllable stereo sound synthesis with minimal data and computational cost. It overcomes the coarseness of traditional semantic descriptors by injecting high-level embeddings at fine temporal resolutions, enabling continuous, granular control over generated sounds while staying within the learned semantic manifold.

At its core, the SCAPES architecture is a small-scale Continuous Normalizing Flow (CNF) [21] that serves as a conditional prior over continuous EnCodec [22] latent representations. Unlike traditional autoregressive architectures that predict sequences token-by-token, the SCAPES architecture is designed to synthesize short audio segments in a single pass. The SCAPES model, however, applies this architecture sequentially over time, conditioning each segment on a memory buffer of previous latent states and high-level semantic embeddings from CLAP [23, 24]. Consequently, the generative process is autoregressive at the segment level, while the underlying architecture remains a non-autoregressive, conditional vector field.

The vector field governing the flow is parameterized by a Transformer [25] and optimized via Flow Matching [26]. Recently, flow-based models have gained significant traction in audio synthesis due to their robust generative capabilities, stable training dynamics, and capacity to model complex continuous distributions without the information loss inherent to discrete, token-based approaches. While these frameworks have been primarily adopted by large-scale, latency-critical text-to-audio systems, SCAPES demonstrates their effectiveness within a lightweight, resource-efficient regime. Furthermore, although not explored in the present work, inference speed can be further accelerated through rectification and distillation processes [27].

This paper is structured as follows: Section 2 establishes the mathematical and methodological background for the proposed architecture and evaluation metrics. Section 3 details the SCAPES model design, encompassing data representation, architecture, and inference strategies. Section 4 presents a comprehensive case study, detailing the dataset, training efficiency, and an evaluation of generation quality, semantic adherence, and long-term stability. Section 5 provides a detailed discussion of the model’s high degree of controllability, its capacity for novel semantic interpolation, and its current limitations. Finally, Section 6 concludes the paper with a summary of contributions and directions for future work.

2. BACKGROUND

This section establishes the foundational mathematical frameworks and methodological concepts underlying the proposed architecture, alongside the specific techniques utilized for generation quality assessment.

2.1. Neural Audio Codecs

Neural Audio Codecs (NACs) are deep learning architectures designed to compress high-fidelity audio into low-dimensional latent

representations for efficient processing, before reconstructing the signal back to the time domain. Recent advancements in NACs generally fall into two paradigms. The first employs a discrete bottleneck, utilizing processes like Residual Vector Quantization (RVQ) to compress latents into discrete token vocabularies, as seen in EnCodec [22] and DAC [28]. The second paradigm seeks to move away from strictly discrete bottlenecks, focusing either on purely continuous representations, as proposed by Music2Latent [29], or hybrid continuous-discrete architectures like CodiCodec [30]. Because the proposed model builds upon the 48 kHz stereo, non-causal variant of EnCodec, we briefly detail its specific formulation here.

In the 48 kHz EnCodec architecture, the input waveform is first normalized, yielding a single scalar factor representing the audio scale. This scale factor is explicitly retained, as it is essential for restoring the signal’s original dynamic range upon reconstruction. The normalized audio is then processed through a stack of convolutional encoder layers, progressively downsampling the temporal sequence into a compact 128-dimensional continuous latent space. It is critical to note that the convolutional operations in this specific implementation are *non-causal*; they rely on a receptive field that requires access to both past and future context to accurately encode the current frame. Typically, these 128-dimensional vectors are subsequently passed through an RVQ module to construct a discrete token sequence. However, when utilizing the codec purely for its continuous representation, this quantization step is bypassed. The unquantized 128-dimensional latents, augmented by their corresponding scale factors, can be fed directly into the non-causal convolutional decoder to be expanded back into a high-fidelity time-domain waveform.

2.2. Continuous Normalizing Flows and Flow Matching

Continuous Normalizing Flows (CNFs) [21] are a class of generative models built upon the framework of Neural Ordinary Differential Equations (Neural ODEs). In this paradigm, rather than relying on a discrete composition of functions as in traditional deep networks, the transformation between distributions is modeled as a continuous-time flow. More precisely, given a time-dependent vector field $F : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$, the generative mapping is governed by the dynamics of the Cauchy initial value problem:

$$\begin{cases} \frac{dx}{dt} &= F(x(t), t) \\ x(0) &= x_0 \end{cases} \quad (1)$$

where the initial condition $x_0 \in \mathbb{R}^d$ serves as the model’s input and is typically sampled from a tractable base distribution (e.g., a standard Gaussian). The generated output is then obtained by evaluating the ODE solution at the final evolution parameter $t = 1$, yielding $x_1 = x(1) \in \mathbb{R}^d$.

The central learning objective in a CNF is to approximate the true vector field using a parameterized neural network F_θ . Flow Matching [26] provides an efficient framework to achieve this, and its mechanism is briefly explained here. First, for any paired sample (x_0, x_1) drawn from the base and target data distributions respectively, one must construct a *conditional vector field* $F(x, t | x_1)$ that explicitly guides the dynamics from x_0 to x_1 . Specifically, integrating this conditional vector field through (1) reveals a deterministic *conditional path* $x_t = x(t | x_0, x_1)$ connecting the two states. It can be shown that the intractable target vector field, which pushes the entire base distribution to the true

data distribution, can be learned by regressing against these conditional fields. This is achieved by optimizing the *Conditional Flow Matching* (CFM) objective:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, x_0, x_1} \|F_\theta(x_t, t) - F(x_t, t | x_1)\|^2, \quad (2)$$

where $t \sim \mathcal{U}(0, 1)$. Minimizing this objective guarantees that the neural vector field F_θ marginalizes over the conditional vector fields, lifting their transport properties globally to recover the complete data distribution (see [26, Theorem 2]).

2.3. Generation Quality Assessment

The Fréchet Audio Distance (FAD) [31] is a standard metric for evaluating generative audio models, measuring the dissimilarity between the mean and covariance of different embedding distributions. However, FAD inherently assumes a multivariate Gaussian distribution, an assumption that often fails for complex data. Moreover, it suffers from finite-sample bias and computationally expensive covariance operations in high dimensions [32]. These limitations can be overcome with the Kernel Audio Distance (KAD) [33]. Based on the Maximum Mean Discrepancy (MMD) framework, KAD is a non-parametric metric that compares distributions in a kernel-induced feature space, capturing complex non-linear statistics without imposing restrictive distributional assumptions.

Both metrics are fundamentally tied to the chosen representation space, and many representations are valid in this context. To ensure a comprehensive evaluation, we use three embedding spaces. First, while acknowledging its statistical limitations, FAD is computed using standard VGGish embeddings [34] to maintain baseline comparability with existing literature. Second, KAD is computed using CLAP embeddings [23, 24] to provide a robust assessment of high-level semantic information. Finally, KAD is applied to perceptual summary statistics derived from TexStat [11], based on the McDermott and Simoncelli texture synthesis framework [9, 10]. By encoding the distributions of cochlear subband envelopes, modulation energies, and cross-channel correlations, TexStat provides a targeted evaluation of the perceptual quality of environmental texture sounds.

3. MODEL DESIGN

SCAPES is designed to synthesize high-fidelity environmental sounds by modeling the distribution of localized latent segments, referred to here as “atoms.” Each atom is generated by a conditional vector field that accounts for a preceding sequence of atoms and a high-level semantic context embedding. This specific data representation is foundational to the model’s architecture and performance. In this section, we detail the formulation of the atomic representation and the rationale behind its design. Subsequently, we describe the architectural framework, focusing on the conditioning mechanisms and the iterative inference strategy. For the sake of generality, we describe all variables abstractly here. For concrete numeric examples see Section 4.

3.1. Data Structure: Atomic Segmentation and Semantic Annotations

SCAPES is agnostic to the specific choice of Neural Audio Codec, provided it utilizes continuous latent representations. This design avoids the “topological shattering” inherent to discrete tokenization, where the acoustic manifold is fragmented into a disjointed

set of indices. By bypassing these rigid structural constraints, the model preserves the natural continuity of the audio signal. For this work, we validate our framework using the continuous latent space of the 48 kHz EnCodec non-causal variant.

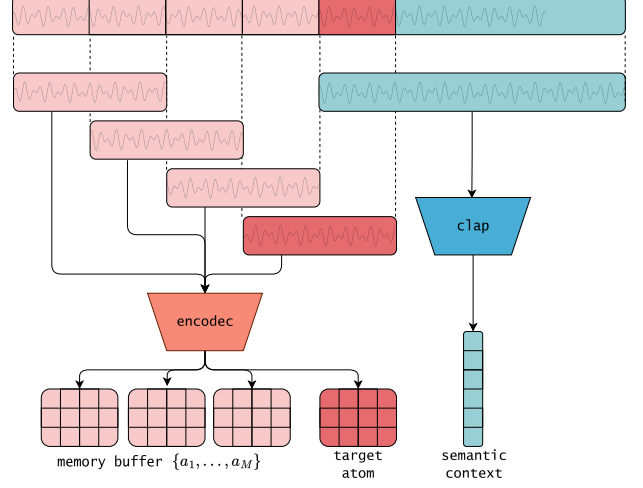


Figure 1: Data representation and annotations. Source audio (top row waveform) is decomposed into overlapping segments (pink and red waveforms) and encoded with EnCodec into *atoms* (pink and red matrices). Each atom is annotated with a semantic embedding (cyan vector) computed from a context window (cyan waveform), extended through randomized repetitions to satisfy CLAP’s minimum input length. During training, the model is tasked to predict the target atom (red matrix) from the preceding atoms in the memory buffer and its corresponding semantic embedding.

Because EnCodec utilizes non-causal convolutions and relies on a global normalization scale factor, naively concatenating independently generated latents produces severe boundary artifacts and phase inconsistencies. To resolve these conflicting receptive fields, we propose an *atomic* segmentation strategy inspired by Saint-Arnaud’s pioneering work on sound textures [35]. We split the continuous audio into overlapping segments and encode each with EnCodec, producing a 128-dimensional latent sequence and a single global scale factor per segment. We combine these into an *atom* $a \in \mathbb{R}^{(128+1) \times N}$, where N is the number of temporal frames. The first 128 dimensions hold the latent features, and the last dimension contains the scale factor, broadcasted along the temporal axis. This unified representation ensures that the latent trajectories and local normalization can be learned together.

While atoms act as a continuous analogue to discrete tokens, they lack the explicit boundary constraints of a fixed vocabulary. In token-based systems, the decoder learns to resolve transitions between discrete indices; in contrast, our atoms are generated independently. To provide contextual continuity, we introduce substantial temporal overlap between successive segments, encouraging the model to capture the joint structure of the sound through redundant acoustic information. This is not trivial because identical acoustic content can correspond to different latent representations, due to each atom’s local normalization and non-causal receptive field. Our experiments show that a Transformer-based architecture effectively reconciles these variations, allowing the model to maintain coherence throughout generation, as further detailed in

Section 4.

Each atom is further paired with a semantic context embedding c , derived from a pre-trained CLAP model [23, 24]. While the CLAP implementation utilized requires 7-second audio windows to produce high-level embeddings, we aim for a finer level of annotation. To achieve this, we extract a context window containing the target atom’s audio, typically shorter than 2 seconds, and extend it to the required 7-second duration via overlapped and added repetition of internal segments of random sizes. This approach generates a continuous, textural extension of the source audio that deliberately avoids introducing artificial rhythmic periodicities, preserves the original timbral characteristics, and provides a robust, localized semantic descriptor.

In summary, this data structure ensures that during training each target atom a_{target} can be conditioned on a preceding sequence of atoms $\{a_1, \dots, a_M\}$, which serve as the model’s temporal memory, along with a corresponding semantic context embedding c . The resulting data pipeline is illustrated in Figure 1.

3.2. Architecture and Flow Matching

The SCAPES architecture consists of a Continuous Normalizing Flow (CNF) where the time-dependent vector field F_θ is parameterized by a Transformer. The primary objective of the model is to learn the conditional distribution of a target atom a_{target} given a sequence of preceding atoms $\{a_1, \dots, a_M\}$ and a semantic context embedding c . This localized generation task allows for a significantly reduced parameter count while maintaining high fidelity, as the model is only responsible for synthesizing short-duration, continuous latent representations rather than long-form waveforms.

The Transformer $F_\theta(x, t | \{a_1, \dots, a_M\}, c)$ processes its inputs through three integrated mechanisms. Within each atom, *self-attention* sub-layers model the local intra-atomic temporal structure using Rotary Positional Embeddings (RoPE) to encode frame-level positions. To ensure coherence across atoms, *cross-attention* sub-layers allow the model to selectively retrieve relevant acoustic context from a memory buffer composed of the M preceding atoms, augmented with a learned atom-position embedding. Finally, global conditioning is incorporated via *Adaptive Layer Normalization Zero* (AdaLN-Zero) [36], where a dedicated MLP modulates the network’s hidden states based on the flow-time t and the semantic context c .

Training is conducted using the Conditional Flow Matching objective, applied independently to the 128 latent dimensions and the scale dimension. We adopt the Optimal Transport (OT) conditional vector field [26]. In this framework, the conditional vector field and paths correspond to

$$F(x, t | x_1) = \frac{x_1 - (1 - \sigma_{\min})x}{1 - (1 - \sigma_{\min})t}, \quad (3)$$

$$x(t | x_0, x_1) = (1 - (1 - \sigma_{\min})t)x_0 + tx_1, \quad (4)$$

where x_1 represents the target atom data, $x_0 \sim \mathcal{N}(0, I)$ is the noise from the base distribution, and $\sigma_{\min}$ is a small variational regularizer used to ensure stability near $t = 1$. This formulation reveals conditional paths corresponding to straight lines and reduces the CFM objective to

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, x_0, x_1} \|F_\theta(x_t, t) - (x_1 - (1 - \sigma_{\min})x_0)\|^2, \quad (5)$$

where, for clarity, the notation suppresses the conditioning on previous atoms $\{a_1, \dots, a_M\}$ and semantic context c in F_θ .

3.3. Inference and Continuous Synthesis

At inference time, the generative process is formulated as an iterative sequence of ODE initial value problems. Each atom is synthesized by solving the Cauchy problem (1) over the interval $t \in [0, 1]$ using a second-order Runge-Kutta method with a fixed number of function evaluations to balance synthesis speed and numerical precision. The vector field F_θ is conditioned on the semantic context c and a sliding memory buffer $\{a_1, \dots, a_M\}$, with Gaussian noise $\mathcal{N}(0, I)$ serving as the initial condition x_0 . This procedure is illustrated in Figure 2. Once a new atom $x_1 = a_{M+1}$ is generated, it is appended to the memory buffer while the oldest atom is discarded, maintaining a consistent temporal context for the subsequent step.

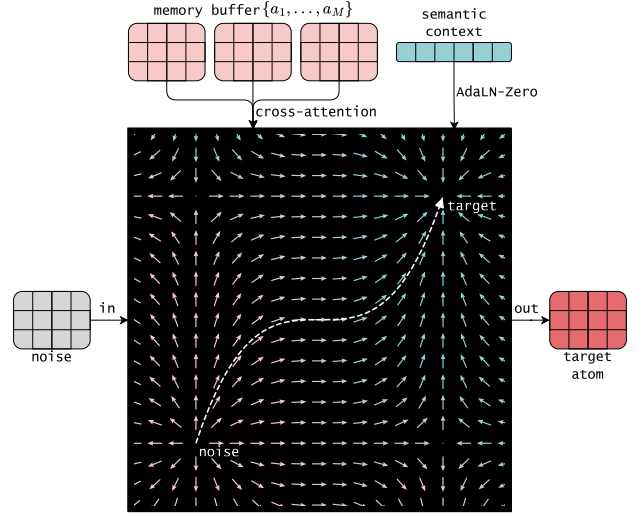


Figure 2: Inference architecture. The Transformer-based vector field F_θ is conditioned on past atomic memory via cross-attention and semantic context via AdaLN. An ODE solver integrates the field from $t = 0$ to $t = 1$ to generate the target atom.

To assemble these atoms into a continuous stream, we utilize a modified Overlap-and-Add (OLA) procedure that builds upon the structural consistency enforced during training. To achieve perceptually transparent transitions, we zero mask the overlapping portion of the new atom, reserving 0.02 s to cross fade between the previous atom tail and the current atom head. This recursive process maintains phase continuity and enables the stable synthesis of coherent, long-form acoustic textures.

4. TRAINING AND EVALUATION

SCAPES has been evaluated across multiple hand-curated datasets with promising results. This section details a specific case study that highlights the model’s data requirements, training efficiency in terms of both time and hardware resources, and the subsequent assessment of generation quality. To support open research and reproducibility, all code, data, pretrained weights, audio examples, and a demo are publicly available on our project page¹.

¹cordutie.github.io/projects/scapes.html.

4.1. Dataset and Training Specifications

The evaluation dataset was curated from Freesound [37] without manual editing and comprises approximately 34 minutes of audio across 10 environmental categories. The sound categories include concert hall applause, crackling bonfires, bubbling water, forest ambiances, a helicopter rotor, keyboard strokes, an orchestra tuning, rain of varying intensities, rivers of different sizes, and wind textures. Following Section 3.1, audio was segmented into atoms of 33 EnCodec frames (0.22 s) with a hop size of 15 frames (0.1 s). This configuration leaves 3 frames of overlap on each side (totaling 0.02 s for cross-fading) and achieves a precise control rate of 10 Hz. Additionally, the memory buffer stores $M = 5$ past atoms, providing 0.5 s of non-redundant audio context.

For this case study, we utilized a SCAPES configuration consisting of approximately 36 million parameters. The model employs 6 transformer layers, each with a hidden dimension of 512, 8 attention heads, and a feed-forward dimension of 2048. Every layer comprises self-attention, cross-attention, and feed-forward sub-layers. Adaptive Layer Normalization (AdaLN) is applied before each, resulting in 18 modulated sub-layers across the network, followed by a final AdaLN module prior to the output projection.

Training was conducted using the Conditional Flow Matching objective for 120 epochs with a batch size of 32 on a single NVIDIA RTX 4090 GPU. The process proved highly efficient, with peak VRAM usage remaining under 8 GB and total training time of approximately one hour. These benchmarks demonstrate that SCAPES can be successfully trained on consumer-grade hardware in a fraction of the time required by massive, token-based models, significantly democratizing high-fidelity sound synthesis for open research. Furthermore, inference is highly efficient: solving the ODE with 16 steps achieves generation speeds twice that of real-time. Nevertheless, to prioritize acoustic fidelity, we employed 32 steps for all experiments detailed in this study.

4.2. Evaluation

In this subsection, we benchmark the generative performance of SCAPES across four complementary dimensions. First, we evaluate objective audio quality by comparing the resynthesis capabilities of SCAPES against RAVE [15], a well-established real-time neural audio synthesis model. Second, we assess semantic adherence by computing FAD and KAD scores to verify that the model maintains class integrity throughout the generative process. Third, we investigate the long-term stability of the autoregressive chain, addressing the "drift" problem typical of iterative latent modeling. Finally, we provide some qualitative observations on semantic interpolation, exploring the model's ability to navigate the continuous latent manifold between disparate acoustic categories.

4.2.1. Generation Quality and Resynthesis

Evaluating SCAPES presents a unique challenge, as its small scale, controllability, stochastic generation, and multi-class capabilities make finding a direct baseline difficult. To facilitate a controlled evaluation of audio quality only, we approximate a resynthesis task by conditioning the model on ground truth past audio and pre-computed CLAP context embeddings. We selected a RAVE model as our benchmark due to its similarly small scale, high fidelity reconstruction, and shared capacity for modeling multiple sound categories within one architecture. We acknowledge that this comparison is inherently asymmetrical: RAVE learns its latent space

from scratch for complete signal reconstruction, whereas SCAPES learns only the data distribution on top of a frozen NAC with massive hidden compute already baked in. Nevertheless, the practical computational differences for a researcher or artist are significant. In our experimental setup, SCAPES operated on less than half the VRAM demanded by RAVE and reached convergence roughly 90 times faster.

To quantify performance, we employed the metrics described in Section 2.3, namely FAD (VGGish), KAD (CLAP), and KAD (TexStat). These scores were computed between the original audio and the corresponding resynthesized outputs produced by each model, and are summarized in Table 1. The results show that, although RAVE demonstrates strong signal reconstruction capabilities, SCAPES generally achieves superior distributional similarity.

We hypothesize that the relatively homogeneous nature of the dataset poses a challenge for RAVE, limiting its ability to generalize effectively. In particular, we observe that RAVE struggles to accurately reconstruct highly noisy sounds. In contrast, SCAPES exhibits more robust performance across such cases. This suggests that the flow-matching objective is effective at capturing the underlying statistical texture of the environment, even in the absence of an explicit reconstruction objective.

Table 1: Quantitative comparison of generation quality across various metrics and models. The best results for each metric (lower is better) are highlighted in **bold**.

Source	FAD (VGGish)		KAD (CLAP)		KAD (TexStat)	
	SCAPES	RAVE	SCAPES	RAVE	SCAPES	RAVE
Applause	4.35	39.06	33.16	47.77	38.1	65.32
Bonfire	3.66	3.65	49.37	37.64	13.28	15.45
Bubbles	5.03	21.05	31.89	47.82	10.9	37.87
Forests	11.05	3.49	61.71	52.56	37.28	16.38
Helicopter	6.9	8.0	22.65	27.8	11.17	28.88
Keyboard	1.05	1.51	55.66	85.68	19.73	29.47
Orchestra	5.14	11.42	17.91	37.62	3.14	9.8
Rain	5.67	13.38	66.85	92.02	15.4	57.16
Rivers	10.17	23.16	76.49	85.14	39.33	58.86
Wind	18.69	23.6	30.36	37.25	23.71	7.19

4.2.2. Semantic Conditioning and Class Adherence

To evaluate the model's adherence to semantic conditioning, we performed fully autoregressive generation for each of the 10 acoustic categories. Each generation was initialized with a "cold start" (zero-initialized memory buffer) and conditioned on the CLAP embeddings derived from the training set. The resulting synthetic audio was then benchmarked against the original dataset using the distributional similarity metrics defined in Section 2.3.

Figure 3 illustrates the resulting similarity matrices across the three evaluation frameworks. In all cases, the matrices exhibit a prominent diagonal structure, demonstrating that in general, generated samples are significantly closer to the reference data of their intended class than to any other category. This behavior confirms that SCAPES successfully captures the distinguishing acoustic features of each environmental texture. Furthermore, the high degree of consistency across VGGish, CLAP, and TexStat representations indicates that this adherence is not merely a reflection of the conditioning space (CLAP), but is also preserved for example in terms of low-level perceptual texture statistics (TexStat).

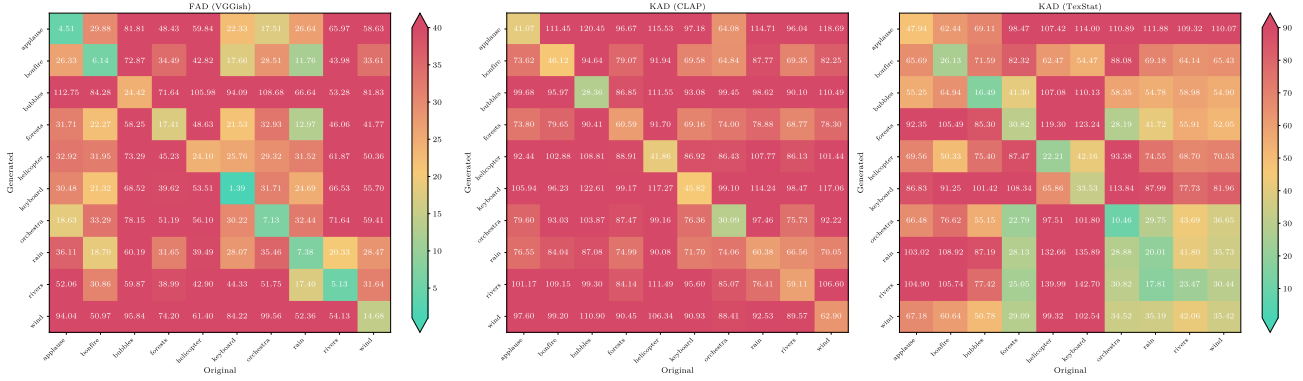


Figure 3: Class-wise distance matrices for generated versus original audio. Lower values indicate greater acoustic similarity. From left to right: FAD (VGGish), KAD (CLAP), and KAD (TexStat), latter two sharing a colorbar. FAD and KAD are not directly comparable. The KAD scores, however, share the same scale due to the use of a common scale invariant kernel, although their interpretations differ.

These results suggest that the model synthesizes high-quality textures that are both semantically faithful and stochastically varied, characteristic of natural environmental sounds.

4.2.3. Long-Term Stability and Autoregressive Drift

To assess long-term stability, we conducted a robustness study across all dataset categories. Autoregressive models are prone to error accumulation, which can lead to "drift" toward out-of-distribution states, feedback loops, or collapse into silence or noise. We generated 125-second audio sequences fully autoregressively from a cold start across 100 trials, using semantic context embeddings for conditioning computed from the dataset. CLAP embeddings were then computed over 5-second segments at 30, 60, and 120 seconds, and compared to the corresponding conditioning embeddings.

As shown in Table 2, a high degree of similarity is maintained over time, indicating stable long-term behavior. These results suggest that SCAPES remains well anchored to the intended semantic manifold, with minimal drift or degradation, and that its conditioning mechanisms operate effectively and in synergy to preserve long-term consistency.

4.2.4. Qualitative Analysis of Semantic Interpolation

A compelling property of the SCAPES architecture is its ability to perform smooth interpolations between different acoustic classes. By traversing the CLAP conditioning space at inference time, utilizing spherical linear interpolation (Slerp) between two target class embeddings, the model can be steered to generate intermediate textures. Moreover, because SCAPES operates on a continuous latent manifold rather than a discrete codebook, these transitions avoid the "quantization jumps" or sonic artifacts typically associated with token-based models.

Preliminary interpolation results are available on the project page and can be explored through our interactive demo. Qualitative inspection indicates that the model supports smooth transitions between acoustically and semantically related soundscapes. For example, interpolating from a creek to a river gradually increases the perceived intensity and density of flowing water, while transitions from gentle rain to a thunderstorm progressively introduce heavier rainfall, stronger wind, and thunder. More surprisingly, the

model also produces plausible intermediate textures between semantically unrelated classes exhibiting similar acoustic characteristics. For instance, an interpolation from keyboard strokes to bonfire crackling gradually dissolves the rhythmic keyboard impulses into stochastic crackling while introducing the low-frequency hum characteristic of a bonfire, ultimately yielding a coherent hybrid environmental texture.

While we acknowledge that the objective evaluation of such hybrid sounds remains an open challenge in generative audio, these results suggest that the learned vector field F_θ is well-regularized and capable of generalizing to regions of the semantic manifold not explicitly represented in the training data. This capability points toward the potential for SCAPES to serve as a tool for fine-grained creative sound design, allowing users to navigate a continuous spectrum of environmental textures.

5. DISCUSSION

While the trend in generative audio continues toward large-scale networks with billions of parameters trained on vast, proprietary datasets, a counter-movement has emerged prioritizing openness, adaptability, and community-driven agency [38, 6]. SCAPES is positioned within this paradigm of "small-scale" AI, offering a fully open-source pipeline that includes the core architecture, curated data, and pre-trained weights. The primary strength of our model lies in its extreme adaptability; because the framework is trained without manual annotations and is computationally efficient, researchers can curate and train on domain-specific datasets in remarkably short timeframes using consumer-grade hardware.

The controllability of SCAPES is another distinguishing factor. The model can operate in several modes: generating indefinite, non-repeating environmental textures from a single semantic embedding; performing "creative continuation" of existing audio; or navigating the semantic manifold via the interpolation strategy detailed in Section 4.2.4. The success of the latter can be attributed to three key factors. First, the conditioning signal is purposefully "relaxed": since the CLAP embedding is computed from a context window that extends beyond the target atom, the model treats the semantic prompt as a directional guide for a transition rather than a rigid, instantaneous constraint. Second, during training, the model learns to map the distinct "islands" of the sparse CLAP manifold

Table 2: Long-term stability analysis. Values represent the mean cosine similarity ($\pm$ standard deviation, where 1 denotes a perfect match) between the target context embedding and 5-second segments of generation after different periods of time.

Source	Cosine Similarity (CLAP)		
	30 s	60 s	120 s
Applause	0.76 ± 0.12	0.71 ± 0.09	0.78 ± 0.09
Bonfire	0.73 ± 0.11	0.73 ± 0.12	0.74 ± 0.12
Bubbles	0.86 ± 0.04	0.89 ± 0.03	0.88 ± 0.04
Forests	0.75 ± 0.06	0.81 ± 0.07	0.77 ± 0.06
Helicopter	0.93 ± 0.02	0.93 ± 0.03	0.86 ± 0.04
Keyboard	0.83 ± 0.06	0.81 ± 0.06	0.81 ± 0.07
Orchestra	0.82 ± 0.07	0.89 ± 0.03	0.81 ± 0.09
Rain	0.87 ± 0.06	0.91 ± 0.03	0.91 ± 0.04
Rivers	0.80 ± 0.08	0.83 ± 0.07	0.81 ± 0.07
Wind	0.87 ± 0.06	0.87 ± 0.06	0.87 ± 0.04

to specific acoustic textures. When traversing the unrepresented space between these clusters, the model must solve a continuity task, synthesizing plausible intermediate states that satisfy both its autoregressive memory and the shifting semantic target. Finally, the use of Optimal Transport Flow Matching [26] yields a highly regularized, continuous vector field. Because the generative process is deterministically governed by smooth ODE trajectories, gradually shifting the semantic condition produces a smooth evolution in the acoustic output, a property only possible in continuous latent spaces.

Despite its versatility, SCAPES possesses inherent limitations. The model currently struggles with signals requiring strict long-term structural dependencies, such as speech or complex polyphonic music. This is a predictable consequence of the "atomic" design; by focusing on a 10Hz control rate and a limited memory buffer, the model lacks the global receptive field necessary to maintain phonemic or harmonic consistency over several seconds. In these contexts, the task of predicting the next 0.1 seconds of audio based purely on local memory and high-level semantics is underspecified. While increasing the memory buffer or incorporating hierarchical conditioning could mitigate these issues, we consider the current architecture to be optimized specifically for the stochastic, textural nature of environmental soundscapes.

6. CONCLUSIONS AND FUTURE WORK

In this work, we presented SCAPES, a Semantically Conditioned Autoregressive Prior for Environmental Sounds. By combining Continuous Normalizing Flows (CNFs), Flow Matching objectives, and the continuous latent representations of EnCodec, SCAPES provides a robust, lightweight framework for high-fidelity environmental sound synthesis. Our evaluation across 10 distinct acoustic categories demonstrates that a 36-million parameter architecture can successfully produce high-quality stereo textures that exhibit strong semantic adherence and long-term autoregressive stability. Furthermore, by operating on a continuous latent manifold, SCAPES functions as an effective interpolation engine, enabling seamless transitions between disparate semantic states, a capability that remains fundamentally challenging for discrete, token-based architectures.

Several avenues for future research remain open. First, while we utilized CLAP embeddings for semantic conditioning, SCAPES

is fundamentally agnostic to the choice of conditioning signal. Future work could explore the integration of low-level acoustic features or human-made annotations. Second, the current model is optimized for a localized receptive field to maintain computational efficiency. Expanding this temporal window through hierarchical attention or sparse transformers could potentially resolve current limitations regarding long-term structural dependencies in speech and music. Third, as the complexity and scale of the training data increase, the scaling laws governing the relationship between model capacity and acoustic diversity warrant further investigation. Finally, while our current inference times are highly competitive, the application of techniques such as rectification and distillation [27] could further accelerate generation, potentially enabling even more complex real-time synthesis applications.

7. ACKNOWLEDGMENTS

This work has been supported by the project "Cátedra en IA y Música" (TSI-100929-2023-1), funded by the Secretaría de Estado de Digitalización e Inteligencia Artificial, the European Union-Next Generation EU funds and BMAT Music Innovators. We also acknowledge support from NVIDIA Corporation and Meta through academic grant programs. Additionally, we would like to express our sincere gratitude to Daniela Quimis, Clara Charbonnier, Jan Pol Obrador, Marcel Manzano, Omar Hamze, Jordi Fabregat, Eduard Herrera, and Eric Mas, for their valuable insights into extending the model to diverse audio sources, exploring alternative training techniques, and helping shape the design of an intuitive demonstration of SCAPES.

8. REFERENCES

- [1] W. W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, no. 1, 1993.
- [2] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," in *International Conference on Learning Representations*, 2020.
- [3] N. Demerlé, P. Esling, G. Doras, and D. Genova, "Combining audio control and style transfer using latent diffusion," in *International Society for Music Information Retrieval Conference*, 2024.
- [4] Z. Novack, K. Saito, Z. Zhong, T. Shibuya, S. Cui, J. McAuley *et al.*, "Flashfoley: Fast interactive sketch2audio generation," in *NeurIPS Workshop: AI for Music*, 2025.
- [5] H. Flores, O. Nieto, J. Salamon, B. Pardo, and P. Seetharaman, "Sketch2sound: Controllable audio generation via time-varying signals and sonic imitations," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2025.
- [6] L. Ibáñez-Martínez, R. Battle-Roca, X. Serra, and M. Rocamora, "MusGU+: Toward a Musician-Centered Evaluation Framework and Discovery Tool for Generative Music AI," in *Conference on AI Music Creativity (Under review)*, 2026.
- [7] K. Karplus and A. Strong, "Digital synthesis of plucked-string and drum timbres," *Computer Music Journal*, vol. 7, no. 2, 1983.

- [8] J. F. O’Brien, C. Shen, and C. M. Gatchalian, “Synthesizing sounds from rigid-body simulations,” in *ACM SIGGRAPH Eurographics Symposium on Computer Animation*, 2002.
- [9] J. H. McDermott and E. P. Simoncelli, “Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis,” *Neuron*, vol. 71, no. 5, 2011.
- [10] J. H. McDermott, M. Schemitsch, and E. P. Simoncelli, “Summary statistics in auditory perception,” *Nature Neuroscience*, vol. 16, no. 4, 2013.
- [11] E. Gutiérrez, F. Font, X. Serra, and L. Wyse, “A statistics-driven differentiable approach for sound texture synthesis and analysis,” in *International Conference on Digital Audio Effects*, 2025.
- [12] A. Barahona-Ríos and T. Collins, “Noisebandnet: Controllable time-varying neural synthesis of sound effects using filterbanks,” *IEEE ACM Trans. Audio, Speech and Language*, vol. 32, 2024.
- [13] L. Wyse and P. T. Ravikumar, “Syntex: parametric audio texture datasets for conditional training of instrumental interfaces,” in *International Conference on New Interfaces for Musical Expression*, 2022.
- [14] A. Dees and S. O’Leary, “Generative latent spaces for neural synthesis of audio textures,” in *International Conference on Digital Audio Effects*, 2025.
- [15] A. Caillon and P. Esling, “Rave: A variational autoencoder for fast and high-quality neural audio synthesis,” *arXiv: 2111.05011*, 2021.
- [16] G. Hadjeres, M. Ferras, K. Koutini, B. Weck, A. Bittar, T. Hummel *et al.*, “Woosh: A sound effects foundation model,” 2026.
- [17] J. Huang, Y. Ren, R. Huang, D. Yang, Z. Ye, C. Zhang *et al.*, “Make-an-audio 2: Temporal-enhanced text-to-audio generation,” *arXiv: 2305.18474*, 2023.
- [18] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic *et al.*, “Audioldm: Text-to-audio generation with latent diffusion models,” in *International Conference on Machine Learning*, 2023.
- [19] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo *et al.*, “Audiobox: Unified audio generation with natural language prompts,” *arXiv: 2312.15821*, 2023.
- [20] R. Valle, R. Badlani, Z. Kong, S.-g. Lee, A. Goel, S. Kim *et al.*, “Fugatto 1: Foundational generative audio transformer opus 1,” in *International Conference on Learning Representations*, 2025.
- [21] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, “Neural ordinary differential equations,” in *Advances in Neural Information Processing Systems*, 2018.
- [22] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *Transactions on Machine Learning Research*, 2023.
- [23] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, “Clap learning audio concepts from natural language supervision,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.
- [24] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, “Large-scale contrastive language-audio pre-training with feature fusion and keyword-to-caption augmentation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez *et al.*, “Attention is all you need,” in *International Conference on Neural Information Processing Systems*, 2017.
- [26] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” in *International Conference on Learning Representations*, 2023.
- [27] X. Liu, C. Gong, and Q. Liu, “Flow straight and fast: Learning to generate and transfer data with rectified flow,” in *International Conference on Learning Representations*, 2023.
- [28] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved rvq-gan,” in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023.
- [29] M. Pasini, S. Lattner, and G. Fazekas, “Music2latent: Consistency autoencoders for latent audio compression,” in *International Society for Music Information Retrieval Conference*, 2024.
- [30] —, “Codicodec: Unifying continuous and discrete compressed representations of audio,” in *International Society for Music Information Retrieval Conference*, 2025.
- [31] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, “Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms,” in *Interspeech*, 2019.
- [32] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, “Adapting frechet audio distance for generative music evaluation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2024.
- [33] Y. Chung, P. Eu, J. Lee, K. Choi, J. Nam, and B. S. Chon, “Kad: No more fad! an effective and efficient evaluation metric for audio generation,” in *ICML Workshop on Machine Learning for Audio*, 2025.
- [34] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore *et al.*, “Cnn architectures for large-scale audio classification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.
- [35] N. Saint-Arnaud, “Classification of sound textures,” Master’s thesis, Massachusetts Institute of Technology, 1995.
- [36] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *IEEE CVF International Conference on Computer Vision*, 2023.
- [37] F. Font, G. Roma, and X. Serra, “Freesound technical demo,” in *ACM International Conference on Multimedia*, 2013.
- [38] R. Batlle-Roca, L. Ibáñez-Martínez, X. Serra, E. Gómez, and M. Rocamora, “MusGO: A community-driven framework for assessing openness in music-generative AI,” in *International Society for Music Information Retrieval Conference*, 2025.