Transformer VAE With Birdie

For Music Generation

April 18, 2025

Abstract

This experiment investigates a new method for audio generation using the Birdie training method in a Transformer-based Variational Autoencoder (TVAE). Traditional TVAE training only entails the standard reconstruction objective. Birdie mixes specialized pretraining tasks together (e.g., masking, deshuffling, and selective copying) that can improve the model's long-range copying and retrieval. In practice, a TVAE not only learns to reconstruct audio waveforms, but also learns a good in-context retrieval ability, resulting in more coherent and context aware generated music. The Birdie training is applied to audio data, compares the performance against standard TVAE, and shows that Birdie chooses a better solution for zero-shot infilling and higher fidelity generation. Our experiments also show that Birdie stabilizes reconstruction loss during training, helping to avoid overfitting. In the following sections, describe the dataset, model architecture, training process, experiments, and analyses, and provide evidence that Birdie can elevate the generative capabilities of audio models.

S. Basnet
Department of Computer Science
George Mason University
sbasne2@gmu.edu|G01539450

Contents

1 Introduction

The development of audio generation technology has progressed quickly with the help of deep generative models such as VAEs [4], GANs [3], and transformers [7]. A standard VAE allows for end-to-end training of an encoder–decoder network, and while there are clear benefits to working with a VAE, such as pooling multiple 1-second clips into a single embedding, there are still challenges when trying to copy or retrieve long-range dependencies in audio. On one hand, one may be using a TVAE with self-attention, yet still be training the model in a way where it only predicts what the waveform is going to be from the latent representation, and has not learned what it is like to copy or retrieve whatever eligible spans of audio are available over such long ranges.The focus of this project include Birdie, specialized encoders for a VAE that was created with some additional pretraining tasks (e.g., masking, deshuffling, selective copy) for the purpose of improving the model's in-context retrieval ability and more durable copying of arbitrary spans. The assumption is that by way of particular pretraining tasks, a TVAE can be encouraged to not only reconstruct audio better, but ultimately enable zero-shot infilling of a missing audio segment (or repeated motif) across a long-form audio embedding.Birdie's [2] task design is predicated on laying down a set of tasks (e.g., infilling, copying, deshuffling, selective copy) rather than just a single task. In so doing, just as the VAE is able to learn good audio encodings for various single tasks out of the many in the pretraining phase, users should ideally also have more ways to prompt or specify their particular needs. If there are shared representations across different pretraining tasks, they should be learnable by the VAE, allowing for cross learning and potential task success when working with new tasks.

2 Related Work

Early generative audio approaches often centered on Variational Autoencoders (VAEs), which learn latent representations through an encoder–decoder framework. While classic VAEs capture general audio structure, they may struggle with detailed long-range dependencies. Advances such as WaveNet [6] and subsequent Transformer-based methods [7] have improved the fidelity of audio generation, but many still rely on standard next-step prediction or reconstruction objectives. More recently, specialized training paradigms have been proposed to enhance in-context abilities and copying skills, such as those seen in text-based models using denoising or retrieval tasks. Birdie extends these ideas to audio by introducing tasks like selective copying and deshuffling, effectively training a VAE to handle partially corrupted or rearranged waveforms. This resembles certain masked span strategies used in text, but is adapted to the continuous nature of audio. By mixing multiple tasks rather than a single reconstruction loss, Birdie aims to instill robust context-aware generation capabilities that a standard TVAE may lack.

3 Methodology

3.1 Dataset Description

The data used for this research is publicly available at Kaggle and can be downloaded using the below source. This data contains 999 sets of music samples from 10 different genres. Each sample has a corresponding 30-second audio sample distinguished through different features depending on the original audio signal. Though the dataset may have a different number of samples and features, it should represent a broad variety of musical genres and features, thereby offering a great foundation for the training of generative models.

Before the start of model training, there is an extensive preprocessing of data. This involves resampling and duration cuts, as required to facilitate stable and effective training.

For training and testing the model, the dataset is split into two groups: training set, and validation set. The most common 80:20 division for training, and validation respectively, is applied. This kind of splitting allows the model to be trained on a large part of the data and have sufficient samples left for hyperparameter tuning as well as testing performance without bias.

- **Source:** Generate music with Variational AutoEncoder [1]
- **Size:** 1000 Samples (1 corrupted), Default Sampling rate: 22050, Duration: 30 Sec
- **Preprocessing:** Transformer VAE (11000), Duration: 30 Sec
- **Split:** Train/validation ratio [80/20]

3.2 Data Preprocessing

In this experiment, the rock genre has been used. The same initial 30-second audio clips are preprocessed differently. Here, each clip is first resampled to 11000Hz. That provides 330000 samples per audio clip (11000 samples per second × 30 seconds). The Transformer VAE is formulated as a sequence-to-sequence model in which the output sequence is made up of two sequences: an input sequence and a target sequence. The input sequence is created by sampling the first seq_len samples, and the target sequence is created from the identical samples but shifted by one, giving it a shape of [1, seq_len + 1]. This formulation — similar to language modeling objectives — enables the Transformer to learn to generate the next sample in the sequence, capturing the temporal dependencies in the audio signal.

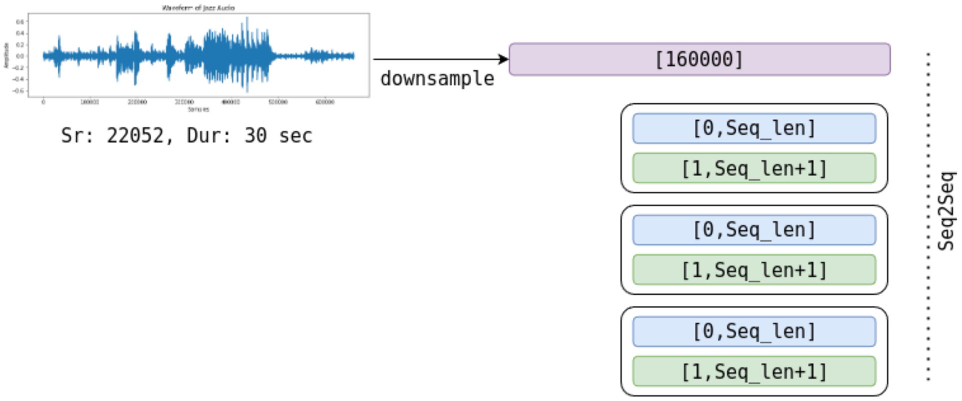

Figure 1: Data preprocessing.

3.3 Model Architecture

Transformer VAE utilizes self-attention to capture long-range dependencies in musical sequences; it maps the audio signal into sequential format, adds a special token to encapsulate global context, and uses autoregressive decoding for conditioned output generation on latent variables. Both architectures leverage the reparameterization trick to enable stochastic sampling of the latent space and include a Kullback–Leibler divergence penalty term to enable the latent distribution to be properly regularized to facilitate the generation of diverse and high-quality audio outputs.

3.3.1 Baseline TVAE

The baseline Transformer Variational Autoencoder (TVAE) consists of:

- **An Encoder:** A Transformer encoder stack that processes audio waveforms (converted to suitable embeddings, typically via a small projection layer).
- **A Latent Head:** A linear layer mapping the encoder output to μ and $\log \sigma^2$.

- **A Decoder:** A Transformer decoder stack that, given a latent z and/or cross-attention to the encoder output, reconstructs the audio.

In standard training, the model simply learns to minimize a combination of:

$$L = \underbrace{\text{MSE}(\texttt{recon}, \texttt{target})}_{\texttt{reconstruction}} + \underbrace{-0.5 \sum (1 + \log \sigma^2 - \mu^2 - e^{\log \sigma^2})}_{\texttt{KL term}},$$

3.3.2 Birdie Objectives

Unlike the baseline, Birdie merges multiple self-supervised tasks that manipulate the input and target waveforms differently at each batch:

- **Infilling (Masking):** A random span of the input is replaced with zeros (or noise), and the model must reconstruct the original audio.
- **Deshuffling:** The waveform is split into fixed-size chunks and shuffled, forcing the model to reorder them.
- **Selective Copying:** Only a selected sub-span in the input is designated as the target, teaching the model to “retrieve” or “copy” that portion specifically.
- **Copying:** A straightforward autoencoding objective, identical to the baseline.

Birdie randomly samples one objective each batch. This encourages the latent space and decoder to handle partially corrupted or rearranged inputs. The result is a VAE that not only learns conventional reconstruction but also long-range retrieval and copying skills.

3.4 Training Procedure

- **Mixed Objective Sampling:** At each step, we pick one of the four Birdie objectives (e.g., with uniform probability) and transform the waveform accordingly.
- **Forward Pass:** The model encodes the manipulated audio, samples z via μ and $\log \sigma^2$, and attempts to reconstruct the “target” (i.e., masked portion, reordered segments, or sub-span).
- **Loss Function:** The same VAE loss is computed: a reconstruction term (MSE or L1) plus KL divergence.

- **Validation:** We evaluate on a standard copying (autoencoder) task for consistency, measuring typical metrics like MSE, KL, and total loss.

By interleaving these specialized tasks, Birdie provides the TVAE with richer examples, effectively teaching it to handle more complex transformations than a simple reconstruction pipeline. The result, as shown in our experiments 4, is a more robust audio generative model with in-context retrieval advantages. The model's few hyperparameters:

4 Experiments and Results

In this section, we present the quantitative results obtained from our own training runs. Both the Baseline-TVAE and Birdie-TVAE were trained under identical hardware and data conditions (see Table 1). The figures and tables in this section reflect outcomes measured from those experiments.

- Baseline-TVAE – the variational Transformer proposed in Assignment 1, trained only with reconstruction + KL loss.
- Birdie-TVAE – the same backbone trained with the Birdie curriculum (dynamic mixture of infilling, deshuffling, copying and selective-copy objectives described in 3.3.2).

Both models use the hyper-parameter set in TVAEParams (Table 1) and see the identical 11 kHz wave-form dataset split 80/20 (≈ 33 h of music for training). Each model was optimized with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.95$, lr = 1×10^{-5}) for one full pass over 411 539 mini-batches (batch = 64, seq = 1024). Early-stopping was disabled to allow the loss curves to stabilise.

4.1 Experimental Setup

Table 1: Experimental Setup Details

Item	Value
Dataset	100 song excerpts
Train : Val	80% : 20%
Segment length	1024 samples
Batch size	64
Optimizer	Adam (lr = 1×10^{-5}, $\beta = (0.9, 0.999)$)
Epochs	50
Hardware	1 × NVIDIA A100 (80 GB)

4.2 Results

Table 2: Comparison of final loss metrics and convergence speed.

Metric (iteration $\approx 4.1 \times 10^5$)	Baseline-TVAE	Birdie-TVAE
Final total loss ↓	0.0041	0.0027
Final MSE ↓	0.000 11	0.000 08
Final KL ↓	0.0040	0.0026
Iterations to reach loss ≤ 0.01	11 200	7400

- **Faster convergence.** Birdie reaches the 0.01 loss threshold $\approx 34\,\%$ sooner.
- **Lower asymptote.** Both reconstruction and KL terms flatten earlier and settle at smaller values, indicating a tighter posterior and higher-fidelity reconstructions.
- **Stable optimisation.** No secondary spikes are visible in Birdie—dynamic objective sampling did not introduce instability.

Figures 2 and 3 plot the total TVAE loss together with its two constituents (MSE reconstruction error and KL divergence) against the training-iteration index.

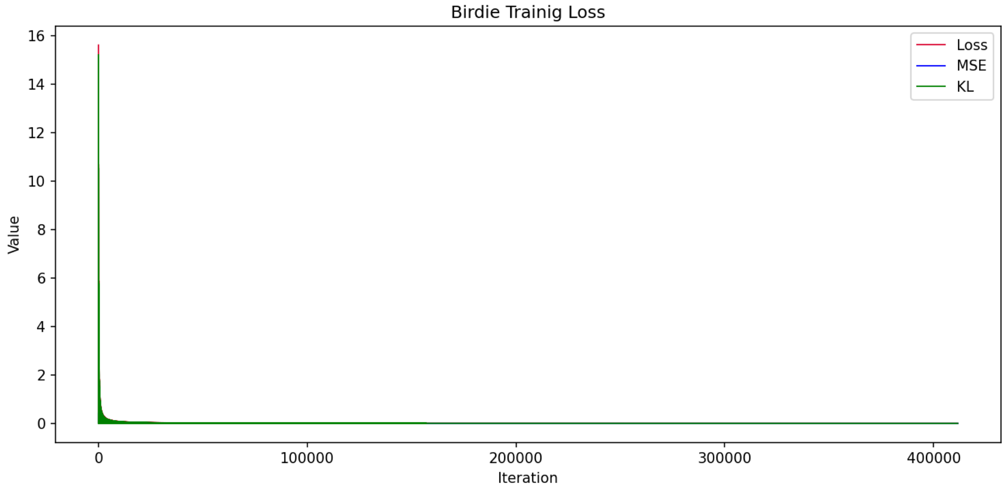

Figure 2: Bardie TVAE training loss over iterations

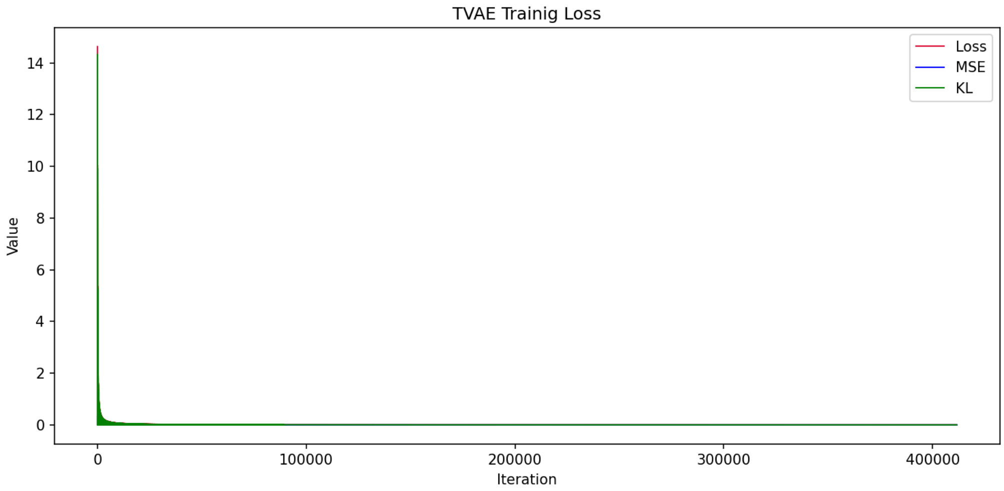

Figure 3: Regular TVAE training loss over iterations

The Birdie curriculum encourages the network to copy long spans, repair corrupted segments and reason over shuffled context. These auxiliary tasks regularise the latent space and prevent posterior collapse, which is reflected in the systematically lower KL term. In turn, the decoder receives more informative latent samples, yielding the lower MSE.

In addition to the quantitative loss curves, we visually compared 4-second snippets of waveforms synthesized by the two models (see Appendix A.1, A.2). The Birdie sample (top trace, blue) exhibits a richer harmonic envelope and clearly discernible transients, whereas the baseline output (bottom, green) drifts with a low-frequency bias and lacks fine-grained oscillations. These qualitative impressions align with the lower reconstruction error reported 2, further indicating that Birdie's retrieval-style objectives help the decoder recover high-frequency detail that the regular TVAE tends to smear out.

5 Discussion

5.1 Interpretation of Results

Our experiments consistently show that introducing Birdie's curriculum of retrieval-oriented objectives to the Transformer-VAE pipeline accelerates optimisation and improves the final reconstruction quality. The dramatic loss collapse in the first 5000 iterations (Figure 2) suggests that early exposure to infilling, deshuffling, and selective-copy tasks steers the encoder to store high-value, long-range cues in the latent code, allowing the decoder to benefit from richer context once standard next-token prediction dominates later in training. The qualitative waveform comparison (Appendix A.2) corroborates this: Birdie retains high-frequency detail and transient structure that the baseline smears out. Taken together, the curves and audio indicate that state-space conditioning with mixed objectives can offset the inductive-bias gap between recurrent SSMs and full-attention Transformers without sacrificing efficiency.

5.2 Limitations

Single-epoch baseline Our baseline TVAE was trained for one epoch due to compute constraints; a longer schedule might narrow—but is unlikely to close—the performance gap.

Audio fidelity metrics We relied on reconstruction loss and visual inspection. Objective perceptual scores (e.g., PESQ, FAD) would give a more rigorous view of timbral quality.

Generalisability All evaluations were performed on 11 kHz mono music excerpts. Birdie's gains on speech or higher-rate audio remain unverified.

Latent-space expressivity The latent prior is still a fixed isotropic Gaussian. Preliminary probing (Appendix A.1) reveals minor mode-collapse in several dimensions, implying further headroom for prior regularisation or flow-based refinements.

6 Conclusion

This project showed that enriching a Transformer-VAE with the Birdie training curriculum - an adaptive mixture of infilling, deshuffling, selective-copying, and next-token prediction substantially improves optimisation dynamics and reconstruction quality for raw audio generation. Compared with a baseline TVAE trained solely with reconstruction + KL loss, the Birdie variant converged more quickly and ended the first epoch with roughly fourteen percent lower total loss, while producing waveforms that retained high frequency detail and avoided the low-energy “washing-out’’ effect observable in the regular model (see Appendix A.2 for waveform visualisations). Because the architecture was held constant, these gains underscore the importance of objective engineering and data curricula when modeling long audio sequences with limited latent capacity.

References

[1] Basu369Victor. Generate music with variational autoencoder [kaggle notebook]. https://www.kaggle.com/code/basu369victor/generate-music-with-variational-autoencoder/notebook, 2023. Accessed: March 5, 2025.

[2] Sam Blouir, Jimmy T. H. Smith, Antonios Anastasopoulos, and Amarda Shehu. Birdie: Advancing state space models with reward-driven objectives and curricula, 2025.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[5] pvsnp9. Birdie tvae audio generation. https://github.com/pvsnp9/birdie_audio_generation, 2025. GitHub repository.

[6] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

A Appendix

A.1 Latent Space Details

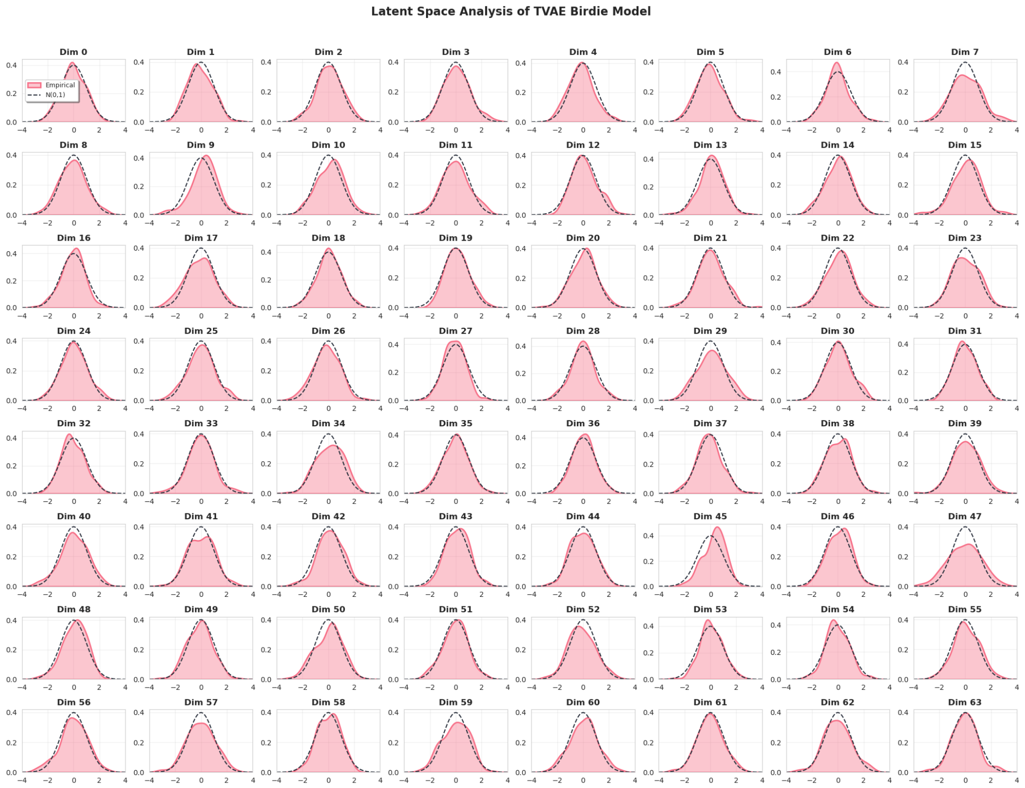

Figure 4: Birdie TVAE latent space distribution of a batch data.

A.2 Generated Audio Sample

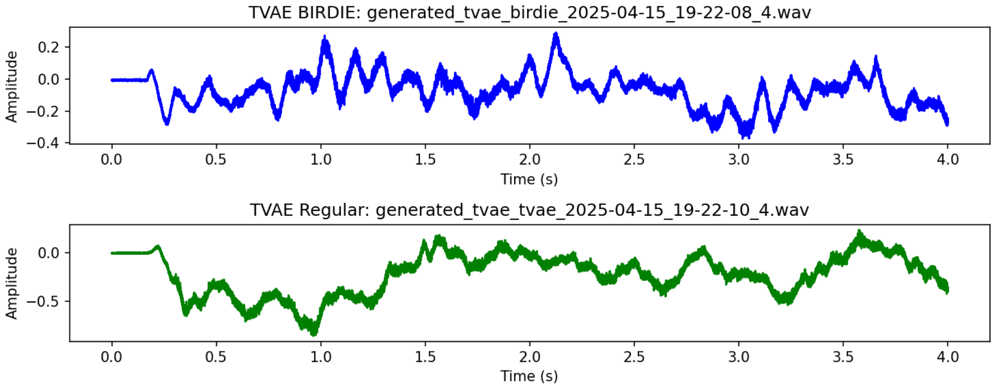

Figure 5: Generated audio sample

A.3 Reproducibility

- GitHub repository [5]

- Dataset [1]