Symbolic Music Generation with Diffusion Models

Gautam Mittal[1], Jesse Engel[2], Curtis (Fjord) Hawthorne[2], Ian Simon[2]

[1]UC Berkeley, [2]Google Research

Code | Supplement

Motivation

Many deep generative models have been used for symbolic music generation

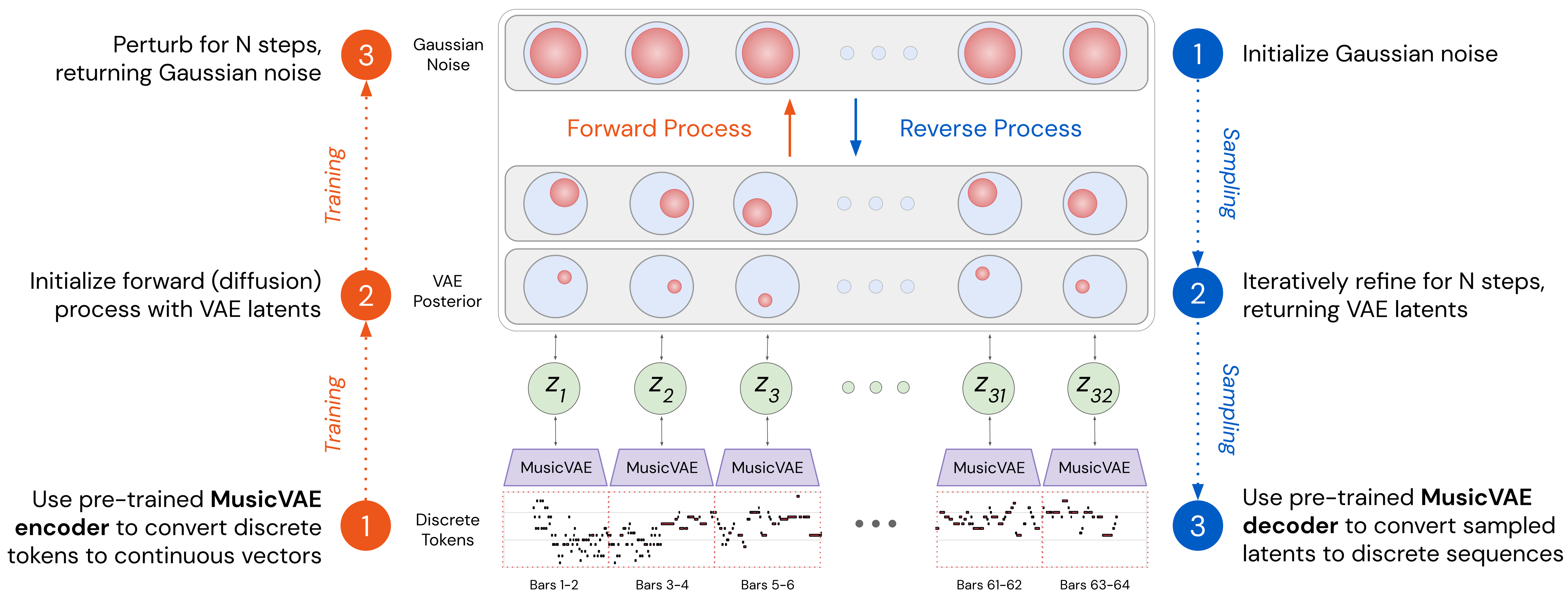
VAE	MusicVAE
GAN	MuseGAN
Autoregressive	Music Transformer
DDPM	***This work***

Denoising diffusion probabilistic models (DDPMs) have produced high-quality results when used to model images, audio, point clouds, etc.

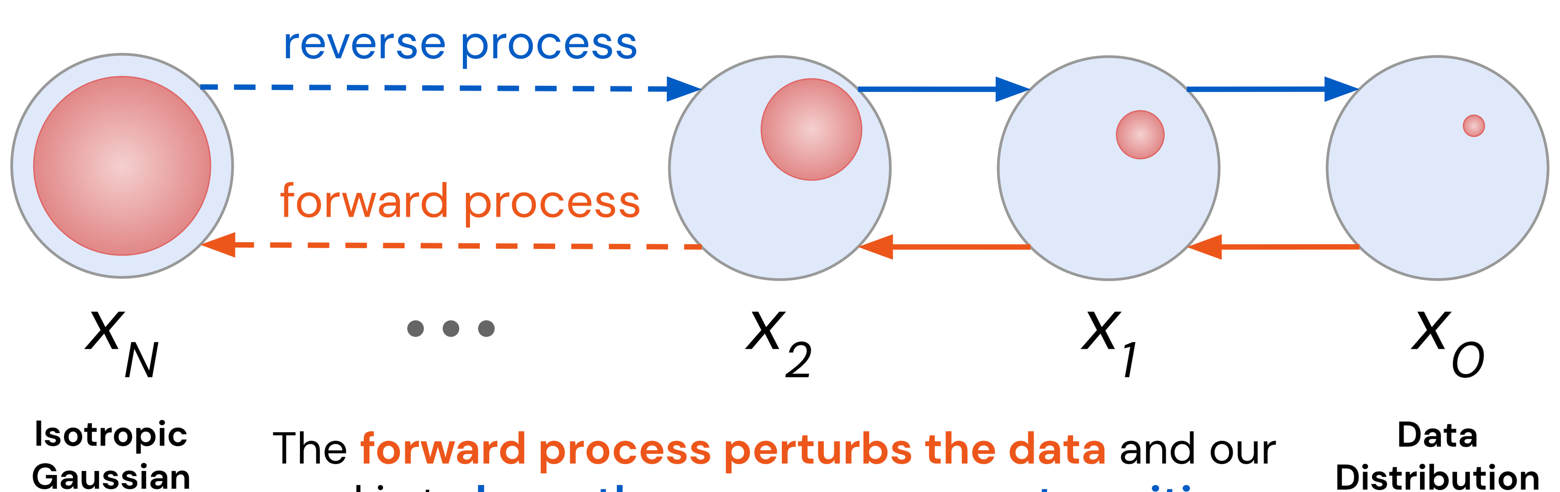
We explore their use for modeling symbolic music data

Model

Key idea: extend DDPMs to music by parameterizing discrete sequences as continuous latent vectors



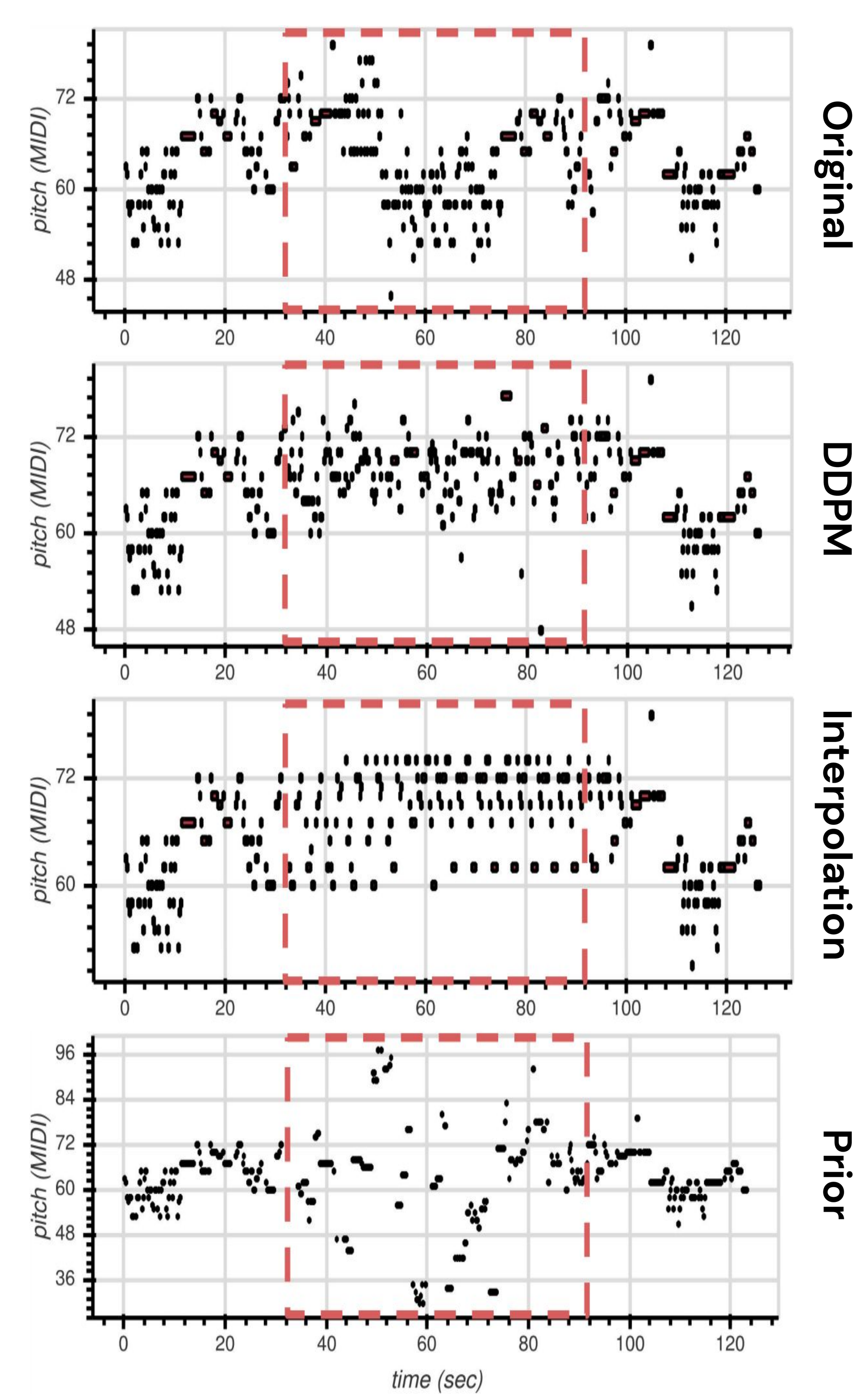
Diffusion Models



The **forward process perturbs the data** and our goal is to **learn the reverse process transitions** to recover the data from Gaussian noise.

Non-autoregressive generation	Latents $x_0, x_1, \ldots, x_N$ are the same dimensionality as the data, allowing parallel generation
Flexible sampling	Iterative refinement steerable generation (e.g. infilling) from a model trained unconditionally
Extension to discrete data	DDPMs assume a continuous distribution while note sequences are from a discrete distribution

Post-hoc Conditional Infilling



Since DDPM sampling uses iterative refinement that is non-autoregressive, we can modify sampling to support infilling without retraining.

We present a note sequence with the middle 32 measures infilled. Our unconditionally trained model produces a plausible result based on the surrounding context.

Future work may explore other ways of modifying sampling for different creative applications.

Evaluation

$$\mathrm{OA}(k, k+1) = 1 - \mathrm{erf}\left(\frac{c-\mu_1}{\sqrt{2\sigma_1^2}}\right) + \mathrm{erf}\left(\frac{c-\mu_2}{\sqrt{2\sigma_2^2}}\right)$$

Framewise self-similarity metrics

$$Consistency = \max(0, 1 - \frac{|\mu_{OA} - \mu_{GT}|}{\mu_{GT}})$$

$$Variance = \max(0, 1 - \frac{|\sigma^2_{OA} - \sigma^2_{GT}|}{\sigma^2_{GT}})$$

Setting	Unconditional				Infilling			
Quantity	Pitch		Duration		Pitch		Duration	
Metric	C	Var	C	Var	C	Var	C	Var
Train Data	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test Data	1.00	0.96	1.00	0.91	1.00	0.96	1.00	0.91
Diffusion	**0.99**	**0.90**	**0.96**	**0.92**	**0.97**	**0.87**	**0.97**	**0.80**
Autoregression	0.93	0.68	0.93	0.76	-	-	-	-
Interpolation	0.85	0.23	0.91	0.34	0.94	0.78	0.96	**0.80**
$\mathcal{N}(0, I)$ Prior	0.84	0.19	0.90	0.67	0.89	0.19	0.94	0.54

Takeaway: DDPMs are promising non-autoregressive models for symbolic music generation and infilling

We'd love to chat! gbm@berkeley.edu